Mining Scholarly Publications for Research Evaluation

Thesis

Mining Scholarly Publications for Research Evaluation

Drahomira Herrmannova

Knowledge Media Institute

The Open University

Submitted in partial fulfilment of the

requirements for the degree of Doctor of Philosophy

30th September, 2017

To my amazing mother and grandmother

Abstract

Scientific research can lead to breakthroughs that revolutionise society by solving long-standing problems. However, investment of public funds into research requires the ability to clearly demonstrate beneficial returns, accountability, and good management. At the same time, with the amount of scholarly literature rapidly expanding, recognising key research that presents the most important contributions to science is becoming increasingly difficult and time-consuming. This creates a need for effective and appropriate research evaluation methods. However, the question of how to evaluate the quality of research outcomes is very difficult to answer and despite decades of research, there is still no standard solution to this problem.

Given this growing need for research evaluation, it is increasingly important to understand how research should be evaluated, and whether the existing methods meet this need. However, the current solutions, which are predominantly based on counting the number of interactions in the scholarly communication network, are insufficient for a number of reasons. In particular, they struggle in capturing many aspects of the academic culture and often significantly lag behind current developments.

This work focuses on the evaluation of research publications and aims at creating new methods which utilise publication content. It studies the concept of research publication quality, methods assessing the performance of new research publication evaluation methods, analyses and ex-

tends the existing methods, and, most importantly, presents a new class of metrics which are based on publication manuscripts. By bridging the fields of research evaluation and text- and data-mining, this work provides tools for analysing the outcomes of research, and for relieving information overload in scholarly publishing.

Acknowledgements

When I was starting my Ph.D., nobody could have prepared me for what was about to come. My Ph.D. journey took me to places I never thought I would visit, allowed me to work on things I never imagined I could work on, and enabled me to talk to people I never thought I would have a chance to meet. Most importantly, thanks to these four years of research, I made friends for life and learned a lot about myself. Looking back, I enjoyed every second of the journey.

First and foremost, I need to thank my Ph.D. advisors and mentors, Petr Knoth and Zdenek Zdrahal. I cannot even begin to explain what having their support throughout these years meant. They always inspired, encouraged, and supported me, they patiently listened to my problems, doubts, and questions whenever I needed help, and always gave me invaluable advice. What's more, they showed me through personal example what it means to be a good researcher. Their faith in me and my work, the opportunities for growth and learning, their advice, guidance, and continuous support had a massive impact on the quality of my research.

I would also like to express a huge thanks to Robert Patton who has been my mentor over the final year of my Ph.D. Robert helped me to gain a very different perspective on my research and inspiration for solving some research problems I have been fighting with for years. I am also very thankful to him for teaching me how to get the message of

my work through and make it stand out, for the countless brainstorming sessions and discussions we had about my work, and for always believing in me.

I must also include a thank you to everyone at KMi for creating such a great space for learning and working. I am proud to have done my Ph.D. at KMi and The Open University, and am thankful to have been given a chance to spend fantastic five years of my life living, studying, and working in England. Thanks to these five years at KMi I was able to make friends with people I hope will stay in my life forever.

I would also like to thank my family. My mom, who is such an incredible role model, was always there for me whenever I needed to talk and has suffered with me through all of my toughest decisions and problems. My grandmother has helped me to become the person I am and passed her love for learning to me. My dad has encouraged my curiosity since a young age and introduced me to technology; I probably wouldn't be in this field if it weren't for him. My grandfather, who was a researcher himself, has always been my inspiration and will forever be in my thoughts, even though we never had a chance to talk. And I cannot forget my brother who has been my support throughout my life in general.

Last but not least, I want to thank my dear Shang who was there for me during the sweat and tears of writing this thesis. Shang, you are the best!

Contents

List of Figures

List of Tables

Part I

Introduction and Background

Chapter 1

Introduction

> *Science can give mankind a better standard of living, better health and a better mental life, if mankind in turn gives science the sympathy and support so essential to its progress.*
> – Vannevar Bush

This thesis deals with the problem of how to evaluate the impact and importance of research publications. Because the amount of scholarly literature is continuously expanding, it is becoming very difficult and time consuming to recognise key research that presents the most important contributions to science. At the same time, given the current economical and political climate, the demand for research evaluation is increasing globally, as there is a clear need to measure scientific progress in order to help fund good research, show returns on investment, and support policy making.

Ever since the first citation index was created [Garfield, 1972], citation analysis has been used to evaluate article impact after publication. Generally, in scholarly publishing, a citation is a reference to a document with the aim of acknowledging influence of the work presented in the document on the publication containing the reference. The most

straightforward and most frequently used way of evaluating article impact is to count the number of times the article has been referenced by other works [Garfield, 1955]. The underlying assumption is that the better the article is, the more people will find it useful and thus reference it in their own work. One can also evaluate the impact of a collection of publications, such as publications written by an author or those appearing in the same journal, by aggregating the number of citations received by that collection. Perhaps the best known indicators, which are based on aggregated citation counts, are the *h-index* for evaluating the impact of authors [Hirsch, 2005] and the *Journal Impact Factor* (JIF) for evaluating the impact of journals [Garfield, 1972] (both of these metrics are discussed in Chapter 2).

However, citations represent only one of many aspects surrounding a publication. Furthermore, the probability of being cited depends on many factors which do not always match the assumption that the better a publication is the more citations it will receive. For example, citations are known to correlate with the number of authors [Bornmann and Leydesdorff, 2015], because more authors can more easily introduce the publication to a wider audience. The way references are used within an article [Shi et al., 2010] as well as free online availability of the article [Antelman, 2004] can influence the number of citations it receives. It has also been shown the more a publication is already cited the more citations it will receive in the future (it receives a cumulative advantage) [Price, 1976]. This effect is reflected in the skewness of the citation distribution [Seglen, 1992].

Furthermore, a significant issue which complicates the development of new measures is the difficulty of assessing the performance of these measures in research evaluation, or, in other words, the difficulty of demonstrating that these measures work and measure some meaningful aspect

of the research process. In fact, the authority of these methods is often established axiomatically. For example, the two metrics mentioned above, the JIF and the h-index, were both proposed without sufficient empirical evidence demonstrating what they measure and how well they work.

To mitigate or avoid some of these limitations, many improvements to the traditional metrics as well as new approaches to research evaluation that do not rely on citation counting have been proposed in recent years. One research strand has focused on mitigating the issues related to the use of citations, for example by normalizing citation indicators by field [Ruiz-Castillo and Waltman, 2015, Colliander, 2015]. The other alternative is to use different data. A number of research studies have investigated utilising data from the Web, such as number of online views or downloads, activity on Twitter, and mainstream media mentions of academic articles or references found in policy documents [Schlögl et al., 2014, Costas et al., 2015, Erdt et al., 2016]. Using web data has several advantages. Online data become available much sooner than citations, which might take years to accumulate depending on the discipline[Glänzel et al., 2003]. These metrics also help to capture broader impacts of research rather than focusing impacts within the research community. However, all these approaches still rely on outside evidence without considering the manuscript of the publication itself.

In this thesis, we investigate how to leverage publication manuscripts in research evaluation with the aim of addressing the above problems. As the idea is to use information that is semantically richer than what has traditionally been used, we call this type of metrics *semantometrics*. In particular, we develop the idea of analysing citation and collaboration patterns in terms of semantic similarity and study how these patterns reflect scientific impact. We approach the question of how to utilise con-

tent in research evaluation methods by breaking it down into a number of steps and sub-questions. We start by analysing the concept of research publication quality to discover the aspects and dimensions of the concept. The discovered dimensions inform the design and focus of our methods. Furthermore, we address the issue of a lack of evaluation data. We do this by identifying typical examples of publications providing high and low volume of change in their particular research area. This way, we are able to compare different metrics based on how well do they distinguish between these types of papers. We then utilise this evaluation method to study the performance of our semantometric measures and show that incorporating content helps to improve the performance of new measures and provides additional information about the quality of research publications complementary to the existing research evaluation measures. To our knowledge, the work in this thesis is among the first to introduce and investigate the use of text analysis in research evaluation. In the following section, we explain the motivations for this work.

1.1 Motivation

Guthrie et al. [2013] have summarised the purposes of research evaluation into four categories: (a) allocation, (b) accountability, (c) advocacy, and (d) analysis. In our view, the purposes of research evaluation presented by Guthrie et al. [2013] are strongly related to the point in the research cycle at which the evaluation is used (to study inputs, outputs, or the research process itself), and we therefore slightly modify the list provided by Guthrie et al. [2013] to match our understanding of the research landscape and merge accountability and advocacy into one category. The three categories can then be summarised as follows:

- **Accountability & advocacy:** advocating or accounting for the

outputs of research, i.e., demonstrating accountability, returns on investment, benefits of supporting research, and good management. Also improving understanding of research among the public and policy makers.

- **Allocation:** determining where to best allocate inputs (funds and resources), i.e., how to distribute funds in order to achieve specific goals.

- **Analysis:** analysing the research process or the outputs at any time during the research cycle to provide support to the research process and to researchers.

This section presents motivating examples related to each of the three purposes of research evaluation.

1.1.1 Accountability & advocacy

There is a need to demonstrate accountability, return on investment, and good management to research funders, taxpayers, and others. In most countries, research and development (R&D) spending constitutes a significant portion of the budget. For example, in 2014 the US spent almost $457 billion on R&D [UNESCO, 2017] (this figure includes both public and private investment and amounts to about 2.7% of US GDP). In the same year the UK spent close to $44 billion on R&D [UNESCO, 2017] (1.7% of the country's budget, the figure again includes both public and private investment). This funding gets distributed to different agencies, institutes, and companies, which in turn distribute their funding to different divisions, groups, and people. Each funding recipient as well as each government ultimately needs to demonstrate the value of the research outputs produced as a result of specific funding, particularly when public funding is concerned. However, due to the complexity

of the academic culture, this a complicated task. For example, metrics typically used to compare countries in terms of their scientific output are number of scientific publications or patents produced by each country, number of scientists employed by a given country, or the number of PhD degrees awarded. While the number of research papers gives an idea of the amount of research done by each country, it omits the quality and significance of the research as well as non-publishable research outputs.

Going back to the example of the US and the UK, in 2014 the US produced over 620 thousand publications, while the publication output of the UK was 180 thousand articles in the same year [Scimago Lab, 2016]. If we consider both public and private R&D spending, this means the cost of one publication was about $737 thousand in the case of the US and $244 thousand in the case of the UK, which is about three times less. However, this doesn't necessarily mean the UK is doing better than the US in terms of research performance. For example, a significant portion of the US federal R&D budget is allocated to defense research [White House Office of Science and Technology Policy, 2014], which, due to its sensitive nature, often cannot be published. The US also funds the largest space agency in the world, the National Aeronautics and Space Administration (NASA); however, the type of research done at NASA might not always result in publications (for example a new space suit design). Furthermore, the US has a long history of commercialising research; however, commercialisation and publishing might in some cases be irreconcilable [Caulfield et al., 2012, Rhoten and Powell, 2007]. Finally, research in the public and private sectors tends to be evaluated differently. While in academia publications are in many fields used as the base unit for evaluation, this may not be the case in the industry where different types of contributions to the company may play a more important role. As a result, countries with a high proportion of private

investment in research may have a lower publication output but still produce high-quality R&D.

Number of patents may give some idea of how well is a country able to turn research ideas into commercial products. However, direct impacts on society like profits and jobs created due to research are much harder to track and are often presented as anecdotal evidence. [Sutherland et al., 2011] summarised three main benefits that research brings to the society:

- Improved life quality or sustainability. This includes research regarding health, the effectiveness of public services, policies, quality of life, or the environment.
- Economical benefits which might come, for example, from linking research with industry and resulting financial profit.
- Contribution to knowledge, in case of research that is driven by curiosity.

Assessing each of these three benefits may require different methods. Many countries, including the UK [Research Excellence Framework, 2014b], the US [Largent and Lane, 2012], and Australia [Australian Research Council, 2015b] have initiated efforts to assess the impact of publicly funded research. These efforts are typically centred around research publications and often require significant manual effort, both for the individuals and institutions being evaluated as well as for the evaluators. This demonstrates the need for automated research publication evaluation methods, which might simplify or completely automate some or all of the related tasks.

However, the evaluation of performance of different countries or institutes is not the only way that researchers might benefit from various quality and impact indicators. The career progression of research

employees is often dependent on how well they can demonstrate their productivity and the quality, importance, and impact of their research [Seglen, 1997, Rossner et al., 2007, Arnold and Fowler, 2011]. Scientific user facilities, such as the Large Hadron Collider (LHC) at CERN or the Spallation Neutron Source (SNS) at the Oak Ridge National Laboratory provide resources to researchers for conducting experiments. With an annual operating cost of about $1 billion for the LHC [Knapp, 2012] and $4 million for the SNS [Department of Energy Office of Science, 2014], the facility managers as well as the funders want to know the impact the facility had [Patton et al., 2012].

1.1.2 Allocation

A different perspective on research evaluation is the perspective related to allocation of research funds and resources, i.e., how to distribute funds in order to achieve specific goals. Internationally, there is a growing interest in utilising science for the technological and economic race and to address societal problems. However, the funds provided by governments for research are often kept tight and focused. For example, while in 1976 public (defence and non-defence) R&D spending in the US constituted over 1.2% of US GDP, in 2014 this number was below 0.8% [American Association for the Advancement of Science, 2017] (that is close to $143 billion [White House Office of Science and Technology Policy, 2014]). Depending on the focus of the standing government, this money is then distributed between several departments including the National Institutes of Health, the Department of Energy, the National Aeronautics and Space Administration, and others.

Another example is a 1993 white paper issued by the UK government, which states that “the decision for Government, when it funds science, as it must, is to judge where to place the balance between the freedom for

researchers to follow their own instincts and curiosity, and the guidance of large sums of public money towards achieving wider benefits, above all the generation of national prosperity and the improvement of the quality of life. [...] The Government does not believe that it is good enough simply to trust the automatic emergence of applicable results which industry then uses." [The U.K. Cabinet Office, 1993]. As a consequence, it is becoming necessary to be able to recognise emerging and growing research topics, centres of research excellence, and scientific experts for funding, hiring, and resource allocation purposes.

Distribution of research funds among research institutes, projects, or people is not the only situation where strategic allocation is needed. Another example is the selection of journal subscriptions. Between 1986 and 2003 the prices of journal subscriptions grew more than three times faster than the consumer price index (CPI) [Panitch and Michalak, 2005] and by 2010 the cost of journal subscriptions grew to almost four times the CPI [Shieber, 2013]. The price growth has reached a point where universities have started announcing they can no longer afford the costs of journal subscriptions [Sample, 2012].

A well-known metric for evaluating journals is the Journal Impact Factor (JIF) [Garfield, 1972]. JIF is based on the number of citations received by the journal and the number of articles published in that journal. Provided that a citation is a demonstration of impact of the cited article, this measure should be sufficient for selecting the most influential journals in a research field. There are, however, many reasons why such metric is not adequate, starting from the simple fact that many journals are cited very infrequently, while some other journals are cited well above average just because of the type of content they publish (for example journals from a very narrow research field vs. review journals) and ending with examples of purposely trying to manipulate and increase

the JIF rating of a journal [Brumback, 2009, Arnold and Fowler, 2011]. In a situation like this the possibility to compare journals based on the quality and importance of research published within them might be of help to institutions.

1.1.3 Analysis

As the amount of research literature is steadily increasing, researchers often rely on various filters to help them reduce the number of articles that they need to read. This is true especially now, when almost all research articles are published online and most research eventually gets published somewhere [Cronin and McKenzie, 1992, Oosterhaven, 2015].

In fact, it was estimated the number of papers published per year across all disciplines to be over 1.5 million in 2008, with over 50 million articles in existence in 2009 [Jinha, 2010]. At the same time, Bornmann and Mutz [2015] have observed the global scientific publication output grows by about 3% each year and the volume of published research doubles about every 24 years. In this environment, it is becoming easier to miss important developments outside of a researcher's domain or potentially influential publications. For this reason, identifying influential and seminal literature is viewed as an important challenge in both research evaluation and information retrieval of scholarly publications.

The current solutions to this problem are typically, as in many other scenarios, based on counting citations. For example, Google Scholar[1], which is one of the major citation indexes, incorporates number of citations in their publication ranking function. In addition it also offers listings of the most cited publications and authors in each area. The Open Access publisher PLOS allows sorting of articles by the number of views and downloads. Another option is subscribing to updates of indi-

[1] `http://scholar.google.com/`

vidual journals which are of interest to the researcher. However, it will be demonstrated that using filters such as these might lead to significant portions of literature being completely ignored.

1.2 Problem statement

1.2.1 Methods for evaluation of research publications

Traditionally, expert peer review has been used as the main filter for controlling both the quantity and the quality of published research, and this method remains the most trusted up to date [Smith, 2006, Nicholas et al., 2015]. The goal of peer review is, as stated by Armstrong [1997] and Nature Neuroscience Editors [1999], ensuring only high-quality works are published or funded. In reality, however, peer-review often fails to recognise false, erroneous, or irreproducible results [Ioannidis, 2014, Begley and Ioannidis, 2015, Teixeira da Silva and Dobránszki, 2015], and there are many known examples of highly cited articles which were retracted due to error or scientific misconduct [Sox and Rennie, 2006, Davis, 2012]. Peer review has also been criticised for often failing to recognise groundbreaking contributions [Campanario and Acedo, 2007, Campanario, 2009] and for reviewer bias, such as due to gender, affiliation, or geographical location [Lee et al., 2013, Walker et al., 2015, Tomkins et al., 2017].

One of the reasons for the issues with peer review is the rapid growth of published research, which was demonstrated in Section 1.1. The more research is published, the more burden it imposes on scientists. This makes it harder for the reviewers to produce a fair review. It may be easier to resort to secondary criteria, such as geographic location, affiliation, or publication record of the authors.

To overcome the issue of the growing amount of literature, many quantitative evaluation methods have been developed over the past decades. The possibly best known and most widely used methods are referred to collectively as *bibliometrics* [Pritchard, 1969]. Bibliometrics include citation-based methods such as citation counting [Garfield, 1955], journal impact factors [Garfield, 1972], h-index [Hirsch, 2005], and similar. The underlying assumption used by these methods is that the better an article is, the more people will find it useful and thus reference it in their own work. These methods have several advantages, mainly their simplicity and accessibility (the JIF is produced yearly by Clarivate Analytics[2], previously by Thomson Reuters, while citation counts received by individual papers can be freely obtained from many online citation indexes, such as Google Scholar[3] or Microsoft Academic[4]).

However, as was mentioned in Section 1, the probability of being cited depends on many factors which do not always match the assumption that the better a publication is the more citations it will receive. A number of researchers have studied the relation between citations and research quality and shown the relation is not clear [Aksnes, 2003, Antonakis et al., 2014, Bornmann and Leydesdorff, 2015]. Furthermore, it has been shown researchers reference papers for a variety of reasons which do not always relate to quality and impact of the referenced research [Nicolaisen, 2007, Bornmann and Daniel, 2008], but instead might be a result of easier accessibility [Antelman, 2004], prior number of citations (this phenomenon is known as the *Matthew effect* of accumulated advantage) [Price, 1976, Seglen, 1992], or prominence of the cited author [Bornmann and Daniel, 2008]. When using citation-based methods, it is important to account for field differences in citation patterns [Brumback, 2009] as well as dif-

[2] `http://clarivate.com/?product=journal-citation-reports`

[3] `https://scholar.google.com/`

[4] `https://academic.microsoft.com/`

ferences between types of research papers [Seglen, 1997]. Furthermore, citation-based methods have been criticised for the skewness of the citation distribution [Seglen, 1992] (according to some researchers, between 55 [Hamilton, 1991] and 90 percent [Meho, 2007] of research remains uncited, while a small proportion of publications receive a high number of citations [Seglen, 1992]) as well as for the ability to purposely manipulate citation counts [Rossner et al., 2007, Arnold and Fowler, 2011]. Finally, a significant drawback of citations is the time they take to start appearing, which, depending on discipline, might be up to several years [Arnold and Fowler, 2011].

Many new methods have been proposed in the past decades with the aim of overcoming these issues. These can be grouped into two main categories. The first category focuses on mitigating the drawbacks of citation counting, for example by excluding certain document types such as reviews from the evaluation [Harzing, 2013] or by normalizing by discipline [Ruiz-Castillo and Waltman, 2015] or number of authors [Van Hooydonk, 1997]. The second group replaces citations with different types of data, particularly data from the Web. The second group includes metrics collectively referred to as *altmetrics* [Piwowar, 2013], which focus on counting online interactions such as social media and news mentions of scientific articles, and so-called *webometrics* [Björneborn and Ingwersen, 2004], which focus on web link analysis. These new Web-based methods offer several advantages compared to the citation-based metrics. For example, while a work's first citation can take years to occur [Brody et al., 2006], online interactions enable tracking the use of a paper often just days after publication [Bornmann, 2014]. However, like bibliometrics, these metrics are based on measuring the number of interactions (although different types of interactions) in the scholarly communication network and are therefore prone to similar issues, such as vulnerability to

manipulation [Bornmann, 2014] and a lack of evidence that they reflect research impact [Thelwall and Kousha, 2015b]. Consequently, none of the new methods have yet become widely used in research evaluation.

1.2.2 From interactions to content

In the previous section we showed that the existing automated approaches to research publication evaluation usually help with reducing the burden of manual research evaluation by counting the number of mentions of a publication, either in other scholarly articles or online. However, most of these approaches face a common problem – **they are fully dependent on external evidence of publication usage**. Nonetheless, as has been discussed in the previous section, assessing the value of a piece of work solely on the number of interactions often does not provide sufficient evidence of quality. Furthermore, the relevance of many publications is recognised only after years or even decades [Van Raan, 2004, Ke et al., 2015], while the majority of publications remains unnoticed, both by other publications [MacRoberts and MacRoberts, 2010] and online [Erdt et al., 2016]. This does not necessarily mean these publications have little value. For example, there are many documented examples of so-called "multiple discovery", a situation where a similar discovery was made by scientists working independently of each other [Troyer, 2001, Whitty, 2017]. Nobel prizes (such as the 2015 Nobel Prize in Physics which was awarded to Takaaki Kajita from University of Tokyo, Japan, and Arthur B. McDonald from Sudbury Neutrino Observatory Institute, Canada, for independently proving neutrino oscillation and that neutrinos have mass) are often awarded to multiple scientists who have independently made a similar discovery. Important discoveries, such as those later awarded with a Nobel Prize, are well documented due to their prominence. However, in many cases, previous similar discoveries might remain unnoticed.

At the same time, a study by Merton [1961] led to the conclusion that "the pattern of independent multiple discoveries in science is in principle the dominant pattern, rather than a subsidiary one". Interaction-based metrics only account for the "discovered" discoveries referenced by others and the only way of identifying "undiscovered" discoveries is by analysing publication content. Thus, considering content is rather important when detecting important and potentially impactful publications.

Furthermore, many of the limitations and drawbacks of the interaction-based metrics can be mitigated or avoided by taking publication content into account. Some of these possibilities were demonstrated in previous work. For example, taking the position [Ding et al., 2013], context [Valenzuela et al., 2015], or sentiment [Teufel et al., 2006] of a citation into account can be used to assign a weight to each citation according to its importance. These approaches have demonstrated utilising content enables incorporating the semantics of citations into evaluation. However, accessing the full content of an article, extracting the plain text, and identifying the context are all notoriously difficult tasks that have been achieved with varying degrees of success [Patton et al., 2012, Klampfl and Kern, 2013, Valenzuela et al., 2015], thus making these existing methods difficult to apply in practice. Secondly, as most publications are never cited [Meho, 2007, Hamilton, 1991] or mentioned online [Erdt et al., 2016], additional metrics which do not rely on these methods are needed.

The work presented in this thesis addresses the question of how to utilise publication content to develop new research evaluation methods which mitigate or remove some of the issues of the existing methods discussed above. In our work, we adopt an approach which is different from the typical methodology used when developing new research metrics. Many works which focus on developing new metrics, particularly those which utilise citations or data from the Web, adopt a data-driven

approach in the sense that they start by collecting and analysing specific data, and only afterwards do they study what the collected data represent. This is a typical approach in bibliometrics, where citation counting has been used since the creation of the first science citation index in the 70s [Garfield, 1972], and where up to this day there is an ongoing discussion about the meaning of citations and whether citations are an appropriate tool for evaluating research [Seglen, 1992, Bornmann and Daniel, 2008, Ricker, 2017]. In contrast to these existing works we start our work by investigating which factors influence research publication quality. We believe an understanding of what constitutes "good" research is important for identifying aspects related to research publications which provide meaningful information. We use this knowledge in our development of new research metrics.

In the rest of this chapter we state the research questions addressed in this thesis, summarise our approach to answering these research questions and contributions made to the state of the art, and provide an outline of the thesis.

1.2.3 Quality, impact, or value?

Before describing our research objectives, we define the basic terminology related to quality and impact of research. The use of the terms "quality" and "impact" in bibliometric research is a common practice. It has been stated that the number of citations a publication receives is a measure of research quality [Bornmann and Haunschild, 2017], as well as that citation counts do not directly relate to quality [Ricker, 2017]. Citation counts have also been used to measure journal impact [Garfield, 1972]. However, no accepted definition of the meaning of these terms in bibliometrics and research evaluation exists. In this thesis we focus on publication *quality*, as in our view, impact (whether it is research, soci-

etal, or other types of impact) is a dimension of quality. As no definition of publication quality exists, we start our research by investigating the concept of research publication quality (Chapter 3, this investigation confirms that impact indeed is one of many dimensions of quality), and the findings from the investigation inform how we think about publication quality. Throughout the rest of this thesis we will focus on publication quality and understand it to mean quality as we define it in Chapter 3. In this thesis we also occasionally use the term "value" when talking about research publications. We define publication value in terms of their quality and use the two terms interchangeably.

1.3 Research Objectives

It can be seen that the area of research evaluation faces a challenge: there is a lack of methods for assessing the value of research publications with sufficient evidence demonstrating these methods measure publication quality. Based on this observation, we formulated the main research question investigated in this thesis as follows:

> **How to effectively incorporate publication content into research evaluation to provide additional evidence of publication quality?**

The main focus is towards providing new methods for assessing the value of research publications by leveraging publication content in a way which will enable applying these methods in practice. Given the limitations of the existing research evaluation methods and issues faced when developing new methods, we have identified the following sub-research questions on which we will focus in the investigation of the main research question:

Question 1: What is research publication quality and what factors influence it?

One of the issues surrounding the existing automated research evaluation metrics is the lack of evidence demonstrating that these metrics provide evidence of publication quality. Although some studies attempted to provide such evidence by investigating the relation between these metrics and peer review [Aksnes and Taxt, 2004, Waltman and Costas, 2014], the methodology used in these studies has been questioned [Aksnes and Taxt, 2004, Ricker, 2017].

Nevertheless, if we wish to measure the quality of research outputs, the first thing we need to do before choosing specific metrics is to discover the dimensions of the concept. Once we have a better understanding of research quality, we can develop methods for assessing some of its dimensions. Therefore, in Chapter 3 we address this question. We start by reviewing the criteria used in different forms of peer review, particularly in journal and conference peer review and in several national evaluation exercises. The rest of the chapter is devoted to presenting the results of a survey which we conducted at the Open University with the aim of gaining a better understanding of the perception of research quality among scientists.

Question 2: How can we evaluate the performance of metrics used in research evaluation for assessing the quality of research publications?

As we have discussed above, the difficulty with validating research evaluation metrics is the lack of evaluation data. A typical data analysis/statistics approach to answering this research question would be to test the metrics on a ranked set of papers and to

express the success rate of these metrics using an evaluation measure such as precision and recall. However, to our knowledge, there exists no ground truth or a reference dataset that could be used for establishing the validity of research evaluation metrics. Because building such golden standard would require significant time and resources we investigate an alternative approach for validating the metrics.

We address this issue in more detail in Chapter 4. We explain the approaches that are typically used for evaluation in this area and build a new dataset which can be used for this purpose.

Question 3: What is the relationship between the existing metrics used in research evaluation and the quality of publications?

Before investigating the possibilities around the use of publication content for evaluation, we examine the existing methods used in research evaluation. We are particularly interested in examining to what extent these metrics capture publication quality and importance and whether these widely used metrics could be improved to capture these publication aspects more accurately. Drawing on our observations and utilising our dataset created in answering the previous research question, we perform an analysis of the existing research evaluation metrics. Furthermore, we study how the existing metrics could be improved without the necessity of incorporating additional data. Through this study we create a new evaluation metric which in our task outperforms the existing methods by a significant margin. The analysis of the existing metrics and our new method are both presented in Chapter 5.

Question 4: How can we use publication content to create new

methods for assessing the quality of research publications?

Using our observations made in answering the previous research questions, we aim to identify and analyse patterns extracted from publication content which could be used to provide evidence of publication quality. We identify a set of interesting patterns that capture the propagation of knowledge between academic publications and between collaborators. Using these patterns we design two new methods which can be used for research publication evaluation. The proposed patterns and evaluation methods are presented in Chapter 6.

Question 5: How can we interpret the performance of the content-based publication evaluation methods and how do these methods compare to the existing metrics used in research evaluation?

Using our dataset developed in RQ2, we study how the patterns and methods proposed in RQ4 help in assessing publication quality. Furthermore, we provide a comparative analysis of these methods with the current research evaluation metrics using a large public collection of documents. The results of this evaluation are presented in Chapter 7.

Finally, to substantiate the research work described in this thesis, our goals are as follows:

Goal 1: Design new methods for assessing the value of research publications and evaluate these methods in comparison with existing research evaluation metrics.

Goal 2: Show that the developed metrics can be deployed in

large document collections to improve the analysis of published research.

1.4 Thesis methodology and outline

Here, we describe the methodology adopted in this thesis. Our research starts with an extensive literature review, presented in Chapter 2. The focus of the review is on identifying and understanding the existing methods and developments in the area of bibliometrics and citation analysis, the existing alternative methods including altmetrics and webometrics, and the methods which utilise publication content. Our review also covers approaches from text-mining which can potentially contribute to this area of work.

To evaluate new solutions for a problem, research evaluation does not typically use the same experimental methodology as other Computer Science tasks, such as evaluation using a ground truth dataset or using human evaluators. One of the reasons for this is the lack of evaluation data which was briefly explained in the previous section. Thus, following the literature review, our methodology starts by investigating what is research publication quality and which methods are typically used for the evaluation of research metrics. This work is described Chapters 3 and 4.

Next, we focus on the analysis of the state-of-the-art metrics for research evaluation identified in our literature review and on finding a way to exploit these metrics to provide new methods for research evaluation (Chapter 5). Following this analysis, we propose a set of patterns extracted from publication content and two new research evaluation methods based on these patterns, which are presented in Chapter 6.

The final part of our methodology described in Chapter 7 is to assess

the validity of our methods. To accomplish this task, we evaluate our methods in two separate studies, one using our dataset developed while answering RQ2, and one using a comparative study with existing research evaluation metrics.

The material of this thesis is distributed in individual parts and chapters as follows:

Part I: Introduction and Background.

Besides the introduction in **Chapter 1**, **Chapter 2** provides the background for our work. We start by defining basic concepts in research evaluation, such as units and levels of evaluation. We then survey the state of the art in research publication evaluation. We categorise the surveyed methods according to their input data into citation-, web-, and text-based.

Part II: Evaluation of Research Publications.

In the second part of this thesis we focus on answering our research questions. Each chapter addresses one research question.

In **Chapter 3** we review the concept of research publication quality and present results of our survey on researchers' perspective of the concept.

In **Chapter 4** we present our analysis of the methods that can be used to analyse the performance of research metrics and introduce our dataset developed for this task.

In **Chapter 5** we present our analysis of the existing research evaluation metrics and a new metric we designed as an improvement over the existing metrics.

In **Chapter 6** we present two new methods for research publication evaluation which incorporate publication content into the eval-

uation.

In **Chapter 7** we further analyse and evaluate our methods introduced in Chapter 6.

Part III: Conclusion.

In **Chapter 8** we discuss the work presented in this thesis, highlight our contributions, present our main conclusions, discuss the limitations of our work, and point out future work.

1.5 Publications

The chapters of this thesis are based on the following publications:

Chapter 4

- Drahomira Herrmannova, Robert M. Patton, Petr Knoth and Christopher G. Stahl. (2017). *Citations and Readership are Poor Indicators of Research Excellence: Introducing TrueImpactDataset, a New Dataset for Validating Research Evaluation Metrics.* In Proceedings of the 1st Workshop on Scholarly Web Mining at the 10th ACM International Conference on Web Search and Data Mining (WSDM), Cambridge, UK. DOI: 10.1145/3057148.3057154.

- Drahomira Herrmannova and Petr Knoth. (2016). *An Analysis of the Microsoft Academic Graph.* D-Lib Magazine, 22, 9/10, Corporation for National Research Initiatives. DOI: 10.1045/september2016-herrmannova.

Chapter 5

- Drahomira Herrmannova, Robert M. Patton, Petr Knoth and Christopher G. Stahl. (2018). *Do Citations and Readership Identify Excellent Publications?* Scientometrics. 115: 239. DOI: 10.1007/s11192-018-2669-y.

- Drahomira Herrmannova and Petr Knoth. (2016). *Simple Yet Effective Methods for Large-Scale Scholarly Publication Ranking: KMi and Mendeley (team BletchleyPark) at WSDM Cup 2016*. In Proceedings of the 2016 WSDM Cup Entity Ranking Challenge Workshop at the 9th ACM International Conference on Web Search and Data Mining (WSDM), San Francisco, CA, USA.

Chapter 6

- Drahomira Herrmannova and Petr Knoth. (2016). *Semantometrics: Towards Fulltext-based Research Evaluation*. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL), Newark, NJ, USA. (**Best Poster Award**). DOI: 10.1145/2910896.2925448.

- Drahomira Herrmannova and Petr Knoth. (2015). *Semantometrics in Coauthorship Networks: Fulltext-based Approach for Analysing Patterns of Research Collaboration*. D-Lib Magazine, 21, 11/12, Corporation for National Research Initiatives. DOI: 10.1045/november2015-herrmannova.

- Drahomira Herrmannova and Petr Knoth. (2015). *Semantometrics: Fulltext-based Measures for Analysing Research Collaboration*. In Proceedings of the 15th International Conference on Scientometrics and Informetrics (ISSI), Istanbul, Turkey.

- Petr Knoth and Drahomira Herrmannova. (2014). *Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution.* D-Lib Magazine, 20, 11/12, Corporation for National Research Initiatives. DOI: 10.1045/november14-knoth.

Chapter 7

- Drahomira Herrmannova and Petr Knoth. (2016). *Towards full-text based research metrics: Exploring semantometrics: Report of Experiments.* Jisc repository, Jisc, Jisc Report 6376.

Additional publications containing aspects of research presented in this thesis

These publications were accepted for publication after examination, but contain aspects of research conducted prior to examination and presented in this thesis:

- Drahomira Herrmannova, Petr Knoth and Robert M. Patton. (2018). *Analyzing Citation-Distance Networks for Evaluating Publication Impact.* 11th Language Resources and Evaluation Conference (LREC), Miyazaki, Japan.

- Drahomira Herrmannova, Petr Knoth, Christopher G. Stahl, Robert M. Patton and Jack C. Wells. (2018). *Research Collaboration Analysis Using Text and Graph Features.* 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Hanoi, Vietnam.

- Drahomira Herrmannova, Petr Knoth, Christopher G. Stahl, Robert M. Patton and Jack C. Wells. (2018). *Text and Graph Based Ap-*

proach for Analyzing Patterns of Research Collaboration: An analysis of the TrueImpactDataset. 1st Workshop on Computational Impact Detection from Text Data (CIDTD) at the 11th Language Resources and Evaluation Conference (LREC), Miyazaki, Japan.

1.6 Invited talks

Beside the conference presentations of the above publications, the following invited talks, discussing the research presented in this thesis, were given during the course of this work:

- **Herrmannova, D.** (2016) *Semantometrics: Towards full-text based research evaluatio.* Jisc Open Citations Workshop, London, UK.
- **Herrmannova, D.** (2016) *Towards full-text based research metrics: exploring Semantometrics.* Jisc Digital Festival Digifest 2016, Birmingham, UK.

Chapter 2

State of the art in research publication evaluation

> *If you want to make an apple pie from scratch, you must first create the universe.*
> – Carl Sagan

In the previous chapter, we have discussed different scenarios which would highly benefit from efficient, effective, and reliable research publication metrics. We have also briefly introduced some of the most common metrics currently used in research publication evaluation and discussed their advantages and disadvantages.

In this chapter, we provide an overview of the literature on research publication evaluation and related areas. In particular, we start by introducing and describing the main elements and dimensions of the research evaluation problem (Section 2.1). This section provides a background for better understanding the literature review. We then review the existing work in the relevant areas (Section 2.2). In Section 2.3 we review the existing initiatives focused on improving research evaluation, including The Metric Tide report [Wilsdon et al., 2015]. In the final Section (2.4),

we provide a discussion of the main strengths, limitations, and gaps in state-of-the-art of research evaluation.

2.1 Background

In this section we provide the reader with the fundamental background knowledge about scholarly publishing and research evaluation, which constitutes the basis of the research presented in this thesis. The aim of this section is to describe the task of research evaluation, describe its purpose and typical uses, and explain which types of research outputs are the focus of this thesis.

2.1.1 Fundamentals

Research evaluation is the task of analysing and evaluating the activities, inputs, and outputs related to scientific research. Research evaluation can be performed using *qualitative* or *quantitative* methods. While quantitative methods are based on predefined metrics that are used to derive information from data, qualitative methods typically involve human judgement which is based on the participants own perception of the studied aspect, and a theoretical interpretation of the results. Qualitative evaluation methods often require extensive data collection such as through expert panels, case studies, surveys or interviews. The UK government's Research Excellence Framework [Research Excellence Framework, 2014b], which is a research evaluation framework for assessing the quality of research at UK higher education institutions (Chapter 3), is an example of qualitative evaluation, while the Scimago Journal & Country Rank [González-Pereira et al., 2010], which is a publicly available portal providing journal and country rankings based on indicators developed from information contained in the Elsevier Scopus database, is

an example of quantitative evaluation. Guthrie et al. [2013] have summarised the purposes of research evaluation, which we have discussed in more detail in Chapter 1. As we have explained in Chapter 1, we view accountability (demonstrating returns on investment) and advocacy (demonstrating benefits of supporting research) as related, and we therefore slightly modify the definition provided by Guthrie et al. [2013] to match our understanding of the research landscape. In our view, the the purposes of research evaluation presented by Guthrie et al. [2013] are related to the point in the research cycle at which the evaluation is used:

- **Allocation:** determining where to best allocate inputs (funds and resources), i.e., how to distribute funds in order to achieve specific goals.
- **Accountability:** advocating or accounting for the outputs, i.e., demonstrating accountability, returns on investment, benefits of supporting research, and good management. Also improving understanding of research among the public and policy makers.
- **Analysis:** analysing the research process or the outputs at any time during the research cycle to provide support to the research process and to researchers.

As Guthrie et al. [2013] pointed out, the choice of tools and methods used in research evaluation will depend on the purpose of the evaluation. A different point of view on the purposes of research evaluation was presented by De Bellis [2009, Chapter Introduction]. This point of view is concerned with the tasks that research evaluation enables:

1. **Information retrieval:** identifying key publications, people, etc.

2. **Research quality control:** measuring the impact of documents and other outputs, as well as journals, authors, and other entities participating in the research process.

3. **Historical and sociological analysis:** study of the history and sociology of science, such as the structure and evolution of scientific disciplines, collaboration between authors, and research fronts and emerging topics.

While the research presented in this thesis may be used at any point in the research life cycle, we are mainly concerned with applications of research evaluation in information retrieval and research quality control.

2.1.2 Evaluation levels

Broadly, there are four levels of granularity at which one typically wants to evaluate impact using article-level metrics as building blocks:

- **Individual publications** and any other types of research outputs (such as measurement data, plots, figures, patents etc.). Methods typically used at this level include citation counting and citation network analysis.

- Journals and conferences or more generally **groups of publications**, for example publications concerned with similar topic or publications created using specific funding sources. Probably the best-known metric used at this level is the Journal Impact Factor.

- **Individual researchers**, who are represented by the set of papers that they published. A well-known metric used to evaluate researchers is the h-index.

- **Groups of researchers**, for example people affiliated with one organisation or institution, country, or other geographic area. Metrics used at this level include for example the Academic Ranking of World Universities (AWRU).

These four levels are highlighted in Figure 2.1. The higher (more general) levels are typically dependent on the lower levels. For instance, the h-index [Hirsch, 2005] can be seen as a generalisation of the traditional citation counts metric to evaluate the impact of researchers. Similarly, techniques to evaluate the importance of publication venues, which are based on information about articles published within them, have also been developed. These include the Journal Impact Factor [Garfield, 1972] and the Eigenfactor [Bergstrom, 2007].

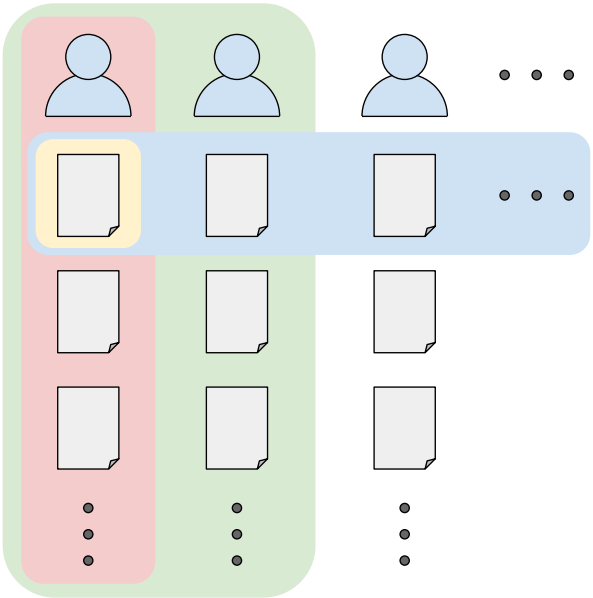

Figure 2.1: Four levels of granularity used in research evaluation. Colour coding is used to highlight different levels: yellow for individual publications, blue for groups of publications, red for individual researchers, and green for groups of researchers.

As the focus of this work is on research publications, we will further

mainly discuss methods related to publications.

2.1.3 Types of scientific publications

As was explained above, our work is concerned with scientific publications. This leads us to the following two questions:

1. What is a scientific publication?

2. What types of scientific publications exist?

Scientific publication is a type of publication, the aim of which is to present and distribute scientific work. The following list summarises the main types of scientific publications. This list was compiled with the help of BibTeX entry types documentation [Patashnik, 1988]. The list is not meant to be exhaustive; rather, the purpose of the list is to provide an overview of the main types of scientific publications, to demonstrate the variety of scientific publications, and to show which publication types are the focus of this thesis. Other types of research outputs include software, data, figures and design; however, these outputs are not discussed in this thesis.

Journal publications are short, in-depth works appearing in online or printed journals or magazines (journals are publications that typically specialise in a particular subject area). In many research areas, journal literature is the most important means of communicating and disseminating research. Journals can (but do not have to) be peer reviewed. Broadly, there are two main types of journal publications:

1. **Primary sources**, such as *research articles*, *letters* (short descriptions of current research intended for fast publication),

and *case studies* (detailed examinations of specific subjects, such as a specific patient). Primary sources describe or report new research.

2. **Secondary sources**, such as *review articles* (articles providing an overview of recent advancement in science, but typically not any original research), *editorials*, *commentaries*, and *letters to the editor*. Reviews may be narrative, or provide quantitative summary of results from the reviewed articles. Reviews are also sometimes called survey articles. These articles typically provide expert opinions, observational studies, comments, discussions, and other analyses of primary sources, but not original research.

Conference proceedings are collections of peer reviewed research papers presented at a conference, symposium, workshop, or other type of meeting. In some research fields, conference proceedings are the main way of communicating and disseminating research (particularly in computer science).

Books are long publications focused on a specific topic. Books are typically written by one or a few authors. The importance and necessity of books varies significantly across disciplines.

Edited books/book chapters represent collections of book chapters, which are typically written by different people and then collected and organised by an editor. Conference proceedings are sometimes published as edited books.

Theses include both Master's and PhD theses. These documents represent authors research and findings conducted in pursuit of an academic degree.

Patents are legal documents which describe an invention (a product or a process) and which provide its owner with exclusive rights to the invention.

Government reports are documents published by a government agency which provide for example details of an investigation.

Project proposals, technical reports and working papers issued either by individual researchers or by organisations. These types of publications typically do not undergo a rigorous peer review. The purpose of these publications can be to present the results of a research project or describe the current state of a problem or project.

Presentations presented at workshops, seminars, or academic conferences.

Online scientific publications including preprints and other research articles published online, for example on a personal web page or in an online self-archiving repository.

Blogs are short articles published in online blogs which might contain opinions and ideas as well as research.

In this thesis we are mainly concerned with journal and conference publications and related types of works, such as preprints. While our methods might be applicable to other types of scientific publications such as books, these may, due to different format, length, and purpose, need a different approach.

2.1.4 Research evaluation sub-disciplines

Before moving on to the literature review, we will describe the main sub-fields and types of metrics applicable to or in some way related to

scientific publications. The names used for these sub-fields are sometimes used interchangeably to describe the whole field. The names relate to types of methods and data being used. Here we provide a brief description of each of these sub-fields.

Scientometrics is a science which is devoted to the study of science and research, or in other words it is a science of science. The term *naukometriya* or scientometrics was created by Nalimov and Mulchenko [1969]. Scientometrics is concerned with scientific productivity, the structure of scientific disciplines, and the relations, history, and evolution of scientific disciplines. Bibliometric indicators are often used in scientometric evaluations, but these are not the only methods and data available – research inputs and outputs (other than publications, for example financial inputs and outputs) and other types of information can also be considered.

Bibliometrics is concerned with any kind of scientific literature or more generally with any kind of written information. The term bibliometrics was first introduced by Pritchard [1969] to describe statistical analysis of recorded information. Pritchard defined bibliometrics as "the application of mathematics and statistical methods to books and other media of communication". The methods used in bibliometrics include counting of articles, books, patents, and other publications, citation analysis, word frequency analysis, and co-word analysis; however, the most frequently used bibliometric method is citation analysis. Bibliometrics is commonly used to asses scholarly impact, but it is also used for other tasks such as studying the evolution of scientific disciplines. Bibliometric methods are also often used in scientometrics; bibliometrics and scientometrics thus overlap to a considerable degree.

Informetrics was according to Hood and Wilson [2001] first proposed in 1979 by Otto Nacke. Simply put, it is a quantitative study of any type of information (including research publications and other outputs). Informetrics applies bibliometric techniques to both scientific and non-scientific publications and written records; it can therefore be viewed as an extension or superset of bibliometrics.

Webometrics takes the informetric methods and models and adapts them for use on the web. The term webometrics was introduced in 1997 by Almind and Ingwersen [1997]. Webometrics is based on the idea that it is possible to view the web as a citation network where nodes are web pages. Björneborn and Ingwersen [2004] divide webometric studies into four main areas: (1) analysis of page content, (2) analysis of link structure, (3) usage analysis, and (4) analysis of web technologies (such as search engine performance).

Cybermetrics has first appeared in a title of a new journal in the same year as webometrics (1997). Cybermetrics and webometrics are related terms which are used to describe the same research area. This allows them to be used interchangeably. Björneborn and Ingwersen [2004] distinguish between the two terms and propose to use webometrics to describe informetric studies of the web and cybermetrics to describe informetric studies of the whole Internet (that means not just web pages and documents but all Internet communication and technology).

Altmetrics is the newest research area of the previously mentioned. The term and the vision of altmetrics (originally alt-metrics, short for alternative metrics) was first introduced by Priem et al. [2010]. The goal of altmetrics is to study science and research by using data from the social web. This includes online bookmarking services,

discussion forums, blog and micro-blog posts, etc. Altmetrics were created as an alternative to the traditional citation counting.

Inspired by Björneborn and Ingwersen [2004], in Figure 2.2 we have attempted to capture the relationships between these fields. The sizes of bubbles in the figure were chosen for clarity, and do not represent sizes of the fields. Because informetrics are defined as a quantitative study of any type of information, the field encompasses all of the other sub-disciplines. Furthermore, scientometrics is defined as a study of all aspects of science, while bibliometrics is concerned with literature (which could be non-scientific), there is therefore a significant overlap between the two fields. Björneborn and Ingwersen [2004] have defined cybermetrics as informetric studies of the whole Internet and webometrics as studies of the Web (predominantly web pages and links between them). Cybermetrics therefore represent a superset of both webometrics and of Altmetrics, which focus on data from the social web (i.e. specific web services instead of web pages).

2.1.5 Terminology

The vocabulary used in research evaluation and scholarly communication research can vary between different publishers, journals, and even authors. Sometimes, one term can be used to describe different concepts, and vice versa. To avoid confusion, in this section we define the terminology which will be used throughout this thesis.

Publication, paper, article: We have defined a scientific publication and listed the main types of scientific publications above. The terms publication, paper, and article are often used interchangeably. The Oxford dictionary definition of the word "publication" is "the preparation and issuing of a book, journal, or piece of music

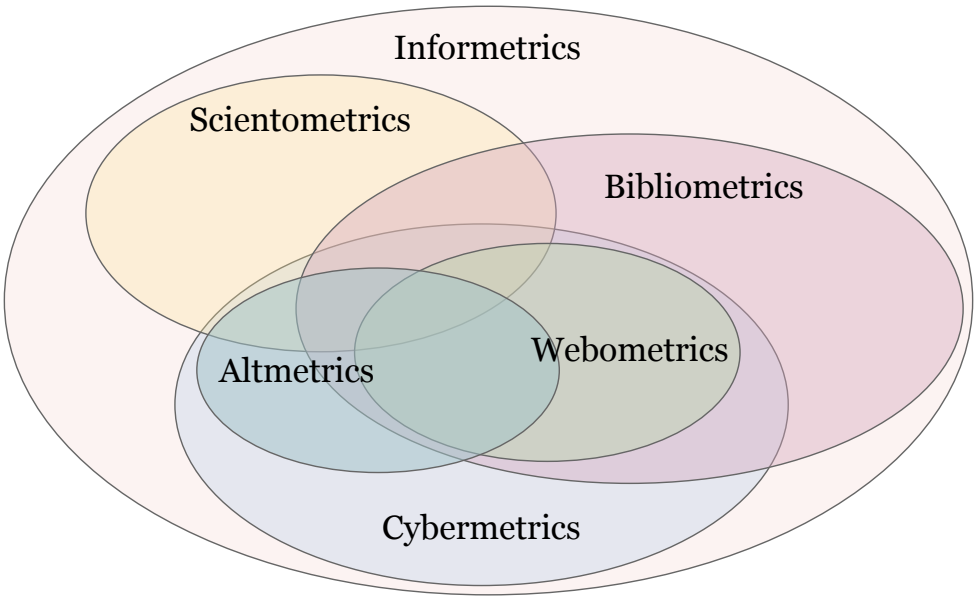

Figure 2.2: Relationships between the fields of scientometrics, bibliometrics, informetrics, cybermetrics, webometrics and altmetrics. The figure is based on a similar figure by Björneborn and Ingwersen [2004].

for public sale". In research, published works are typically peer-reviewed and shared through a journal or a conference; however, the publishing of a work can also be done using a less-traditional method such as online self-publication. A research "paper" is typically understood to be a formal scientific publication describing or presenting research and containing references to other works, while the term "article" is often used with the same meaning as the term "paper". In this thesis, we will also use the three terms interchangeably and understand them to mean formal published works (regardless of the publication method) presenting scientific research and containing references to other works.

Reference: In academic publishing, a reference is a bibliographic description of a research work that identifies the research work. In scientific publications, references are typically listed at the end of

the publication in a section called "Bibliography" or "References".

Citation: A citation is an in-text mention of another (published or unpublished) work, which is typically done using a key referring to a reference found in the publication's bibliography. In this thesis, we will use the term "citation" to mean a link between two scientific publications, the phrase "citing publication" to refer to the publication doing the citing (the publication containing the in-text citation), and the phrase "cited publication" to refer to the publication being cited (the publication listed in the reference section of the citing publication).

Metric, measure, indicator: These three terms are often used interchangeably, although there is a subtle difference in their meaning. *Measure* is typically used to mean a value that is quantified against some standard, *metric* is a way of expressing the degree to which a subject conveys what is being measured, and an *indicator* is a measurement performed against a baseline [Mullins, 2009]. In this thesis we will use these three terms interchangeably and understand them to mean what metric means in the definition above, i.e. a way of expressing the degree to which a subject conveys what is being measured.

2.2 Evaluation of research publications

This section presents the history and the main methods and developments in the field of research publication evaluation. Over the past few decades, scientific publishing has witnessed several important changes – the computerisation of scholarly literature and later the transition of the literature (and the whole publishing process) to the Internet, the creation

of the first citation index of scholarly literature, and the birth and growth of Open Access publishing. Each of these changes reflects in the evolution of this research field. Here, we review the history and the recent developments in bibliometrics, particularly those developments which are in some way related to research publications (we do not review topics such as patent analysis and visualisations of bibliometric data, as those topics are out of the scope of this thesis).

There is a significant overlap between scientometrics, informetrics, and bibliometrics, as bibliometric methods are used to evaluate and analyse research publications in both scientometrics and informetrics. For this reason, out of the three fields, we focus on bibliometrics. We also review the recent developments in webometrics and altmetrics, as both fields are relevant to evaluation of research publications. This review is not meant to be comprehensive, but aims at providing an overview of the main developments, directions, and concepts in these fields.

We categorise the methods reviewed in this chapter somewhat differently than usual. Our categorisation revolves around the input data used by the different methods. There are three main types of data typically used in evaluation of research publications: citations, data from the Web, and text (publication content). In our review we follow this categorisation and focus separately on citation-based methods, web-based methods, and text-based methods. As a result, certain text-based methods (i.e. co-word analysis methods), which are typically categorised as bibliometrics, are reviewed together with other text-based methods (such as automated citation context classification methods) rather than with bibliometrics. This might seem counter-intuitive; however, in our view, this is a more logical organisation of the existing research because the choice of data may influence the capabilities and limitations of the method. First, in Section 2.2.1, we review the history of the field of bibliometrics. These

developments are important because they have shaped the discipline, and many of the original methods are still used today. In Section 2.2.2 we review the current developments in the field of bibliometrics with focus on citation-based bibliometric methods, and in Section 2.2.3 we survey two sub-fields which make use of data from the Web, webometrics and altmetrics. Finally, in Section 2.2.4 we look at the existing methods in bibliometrics and related fields which utilise text for evaluation of research publications. We provide a summary of our findings and conclude the chapter in Section 2.4.

We would like to note that the focus of this chapter is on *methods* for evaluation of research publications. In Chapter 3 we provide a separate review of the concept of publication quality: we review the relevant literature and provide results of a survey we conducted at the Open University on researchers' perspective of publication quality. In Chapter 4 we focus on datasets and methods for evaluating the validity and performance of research metrics.

2.2.1 Foundations of bibliometrics

As mentioned earlier, the term bibliometrics was first introduced in [Pritchard, 1969]; however, bibliometric methods existed and were used decades earlier. The bibliometric study by Cole and Eales [1917], in which the authors examined the amount of literature published in each European country, is often regarded to be one of the first bibliometric studies [De Bellis, 2009].

Bibliometric laws

During the 1920s and 1930s, three important bibliometric studies were published which revealed some important patterns. These studies later became known as the *bibliometric laws*. In 1926, Lotka [1926] observed

that the distribution of productivity among scientists is very skewed, so he created a formula now known as *Lotka's law*. Lotka observed that the number of authors making n contributions is about $\frac{1}{n^2}$ of those making one, and that the proportion of authors making a single contribution is about 60%. That means that approximately 60% of all authors will have one publication, $\frac{1}{2^2} \cdot 0.6 = 15\%$ will have two, etc. Lotka's distribution is shown in Figure 2.3.

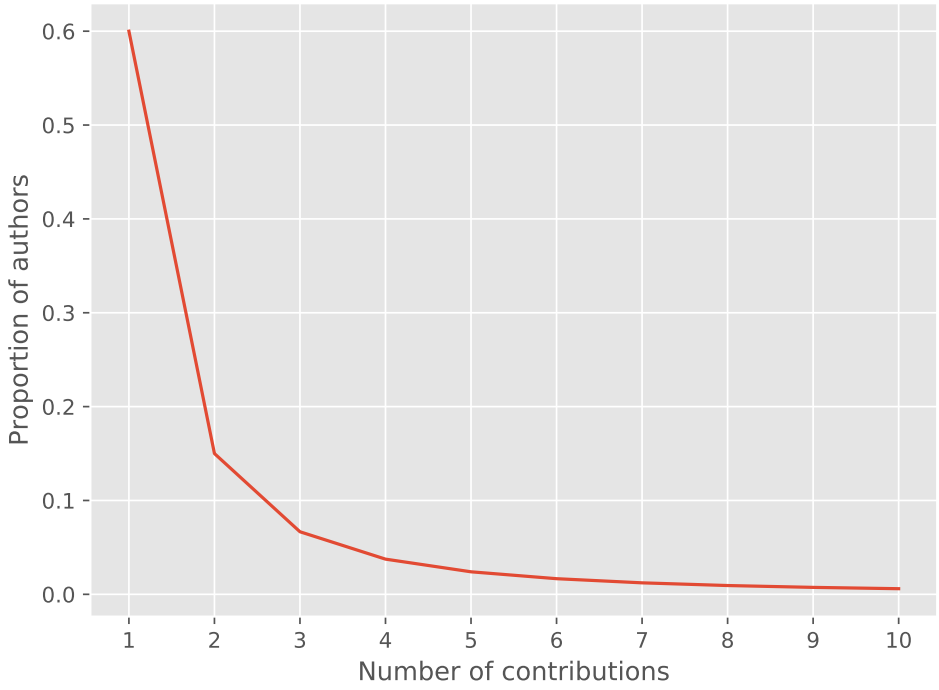

Figure 2.3: Lotka's frequency distribution of scientific productivity.

Later, in 1934, Bradford [1934] first described a pattern of scattering of literature over different journals, which is now called *Bradford's law*. Bradford observed that while some journals are very productive and publish many articles, many more journals are moderately productive and publish far fewer articles. Bradford's law describes this observation. Bradford stated that if all journals were sorted by the number of articles they published from the most to the least productive, and

then divided into three groups with each group containing approximately the same number of publications, the proportion of journals in these three groups would be $1 : n : n^2$. This observation may be useful when managing journal subscriptions, building academic search engines, or collecting data for studies.

Zipf's law was originally used to demonstrate the distribution of words in English text, but it has also been used to model the distribution of citations to academic papers. Zipf observed that when words were sorted by their frequency from the most frequent to the least frequent, their rank was inversely proportional to their frequency [Zipf, 1935]. This can be formulated as $r_i \approx \frac{1}{i}$. The equation states that a word with rank r_i will have a frequency of approximately $\frac{1}{i}$. This relation is depicted in Figure 2.4.

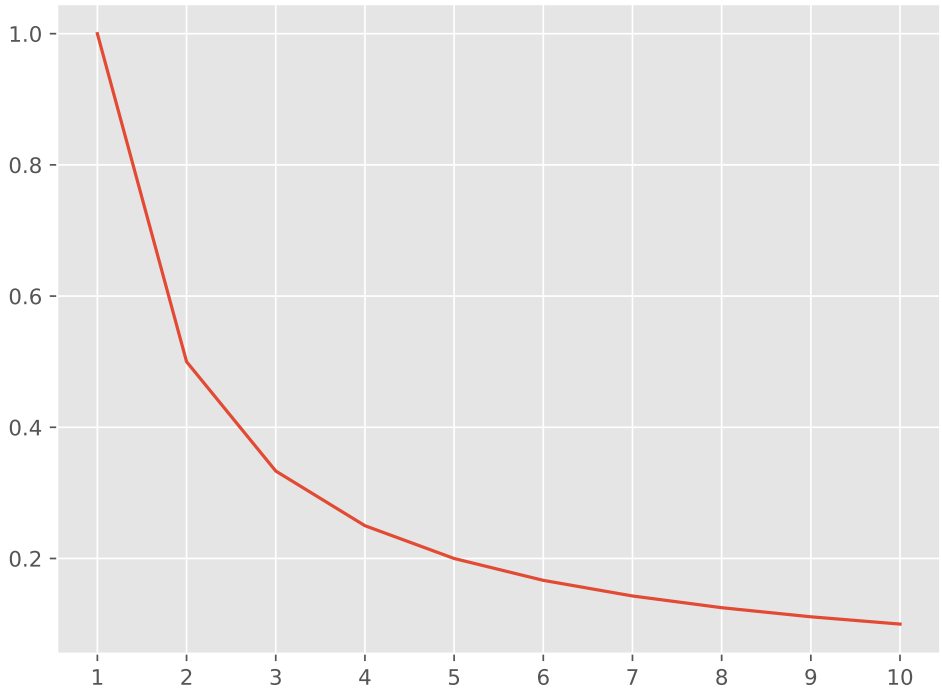

Figure 2.4: Zipf's law distribution.

These three laws can be used to describe many datasets, and similar

distributions have also been found within citation networks [Price, 1976, Seglen, 1992] and on the Web [Björneborn and Ingwersen, 2004]. However, as noted by Brody et al. [2006] these skewed distributions in many cases complicate the analysis of the data, as most statistical methods are based on a Gaussian distribution. Zipf's and Lotka's laws are relevant to our work as in this thesis we work with citation and collaboration networks which conform to these laws.

The Science Citation Index

A major event which helped to speed up the growth and popularisation of bibliometrics was the creation of the first citation index for science. The idea of a citation index for science was put forward by Garfield [1955] and the index came to existence in the 1960s as the Institute for Scientific Information (ISI) Science Citation Index (SCI) [Garfield et al., 1964, Garfield, 2006]. The SCI enabled, among other things, the creation of the Journal Impact Factor, which eventually became a standard for evaluating journals. The SCI is now owned by Clarivate Analytics and is made available through different platforms, such as the Web of Science[1].

Garfield [1955] suggested that a citation index of scientific literature may help to cope with information overload – it simplifies finding relevant literature by tracing citations, improves communication between scientists, and enables them to see the consequences of their work [Garfield, 1955]. In the same paper, Garfield also suggested to use the citation index to measure the impact of published work – the "impact factor" [Garfield, 1955].

[1] `http://www.webofknowledge.com/`

The journal impact factor

The SCI enabled the development of new research metrics including the Journal Impact Factor (JIF) and related statistics such as the *cited half-life* and the *immediacy index*. The idea of JIF was first presented in 1972 [Garfield, 1972]. Since the 1970s, the JIF has been published yearly through the Journal Citation Reports (JCR). The formula for calculating the JIF is as follows:

$$JIF_x = \frac{Y}{Z} \tag{2.1}$$

where JIF_x is the JIF of a journal in year x, Y is the number of citations from articles published in year x to articles published in that journal in years $x-1$ and $x-2$; and Z is the number of "citable items" [McVeigh and Mann, 2009] published in the journal in the years $x-1$ and $x-2$. For example, the 2013 JIF is calculated using data from years 2012 and 2011. In other words, JIF is the mean number of citations received by articles published in a journal during a given time period. The two year window was selected by Garfield [1972] based on the distribution of age of citations to articles, which has shown that typical article is cited most heavily during the first two years after it is published. The JCR also contains a five-year impact factor.

Cited half-life is the median age of articles in a journal that were cited in a selected year [Clarivate Analytics, 2017a], and the immediacy index is the frequency of citations that one article received within specific time period, typically during the year in which the article was published [Clarivate Analytics, 2017b]. Together with JIF, cited half-life and immediacy index form the basis of the SCI metrics [Amin and Mabe, 2004]. While the cited half-life provides a context for understanding how fast publications in a given journal age, the immediacy index gives an

indication of how fast are papers in a journal typically cited.

JIF has become the standard metric for evaluating the "impact" of journals, and it has also been used to evaluate researchers by utilising the JIF of venues in which the researchers published. However, it has been shown the use of JIF especially for the latter purpose is inaccurate for a number of reasons, including the fact citations to articles in a journal follow a similar distribution to those described by Lotka and Zipf, which means JIF does not accurately capture the impact of articles published within the journal [Seglen, 1992, 1994, 1997]. The JIF also does not account for differences between different scientific disciplines, which are in some cases quite significant [Waltman, 2016]. Moreover, some researchers have reported that they were not able to replicate the JIF calculation results [Rossner et al., 2007] or the process of selecting citable items for the JIF calculation [The PLoS Medicine Editors, 2006]. It has also been shown the JIF is susceptible to "gaming" (attempting to increase the number artificially) [The PLoS Medicine Editors, 2006, Rossner et al., 2007, Brumback, 2009, Arnold and Fowler, 2011].

Citation network analysis

Beside of the development of new metrics, the creation of the SCI allowed the analysis of citation networks. One of the first such studies was done by Price [1965] [De Bellis, 2009], who developed models to represent the distribution of citations received by a paper and used this distribution to describe the "active research front" in science. Since then, countless researchers have applied bibliometric and other methods to citation networks in order to analyse the impact and importance of scientific publications and researchers [Waltman, 2016]. Price was also among the first to use citation networks to characterise the growth of science [Price, 1986], to study the cumulative advantage phenomenon [Price, 1976], and to

study the differences in the citation distribution of different fields [Price, 1970].

The availability of citation networks has also fuelled the creation of measures of correlation based on specific citation patterns. These measures include *bibliographic coupling* (two documents are "coupled" if they contain the same reference or references) [Kessler, 1963] and *co-citation analysis* (two documents are co-cited if they are referenced by the same document) [Small, 1973]. Bibliographic coupling and co-citation are depicted in Figure 2.5. Particularly, co-citation is an interesting measure of correlation or closeness because it captures the fact that other authors perceive selected work as similar or related. A similar method has been applied to measure the similarity of authors [White and Griffith, 1981] as well as journals [McCain, 1991].

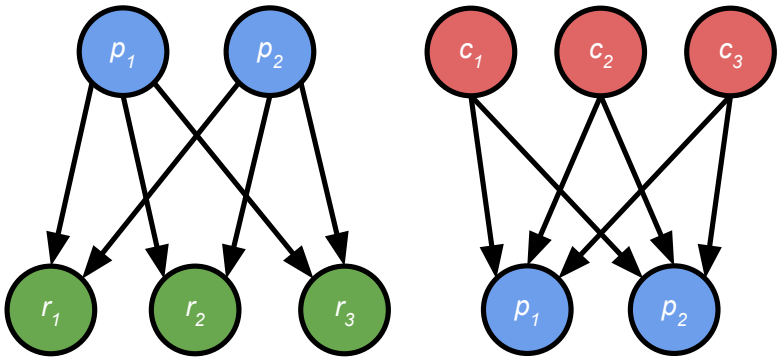

Figure 2.5: A visual representation of bibliographic coupling (left) and co-citation (right). Publications p_1 and p_2 (left) are coupled, because they contain the same references r_1, r_2, and r_3. Publications p_1 and p_2 (right) are co-cited, because they are cited by the same publications c_1, c_2, and c_3.

2.2.2 Bibliometrics today

Since the Science Citation Index was developed in the 1960s, bibliometrics has significantly grown and a staggering number of new metrics and indices have become available, some of which have become very popular among scientists (one such metric is the h-index, which we describe below). In this section we review the current main directions and developments in the field, with focus on bibliometric indicators and methods related to evaluation of scientific publications, and indicators derived from these methods.

Evaluation of researchers

There has been much interest in finding new methods for evaluating individual researchers. One of the reasons has been the increasing use of JIF for evaluating researchers, which has been criticised by a number of researchers [Seglen, 1997, The PLoS Medicine Editors, 2006, Brumback, 2009]. Probably the best known metric for the evaluation of researchers is the *h-index* [Hirsch, 2005]. The h-index is defined as follows: a researcher will be assigned the value h if h of his or her publications have each received $\geq h$ citations and all the remaining publications have received $\leq h$ citations [Hirsch, 2005]. H-index thus captures the number of core highly cited publications of a researcher [Hirsch, 2005]. A similar method can be applied to any collection of research publications and has been applied for example to journals [Braun et al., 2006].

This metric has several advantages over the JIF when used for evaluation of researchers. It is mathematically very simple and captures productivity as well as citation impact. However, it also has some limitations; for example it is field dependent, disadvantages younger researchers, and does not take into account actual number of citations (which

means two researchers with very different number of citations can have the same value of h) [Rousseau, 2008]. For these reasons several variants and replacements of the h-index have been proposed. For example, the *m quotient* divides h-index by the number of years that the scientist has been active (which helps to create a fairer comparison between junior and senior researchers) [Hirsch, 2005], the *c-index* [Bras-Amorós et al., 2010] weights citations by the collaboration distance between authors, and the *g-index* is calculated as the highest number g of papers that receive in total at least g^2 citations (which gives more weight to highly cited publications) [Egghe, 2006]. A similar metric to the g-index is the *a-index* which is the mean number of citations received by the publications in the *Hirsch core* (publications which have at least h citations) [Bihui et al., 2007].

Several recent publications have provided a review of metrics for evaluating researchers and comparisons of the existing metrics on different datasets [Aoun et al., 2013, Díaz et al., 2016, Oberesch and Groppe, 2017]. Yan et al. [2016] have also compared several variants of the h-index for evaluating individual publications. However, despite many developments and new metrics with various advantages and strengths, the h-index remains the most popular metric among the research community (possibly because it is readily available in the largest search engine for academic literature, Google Scholar).

Allocating credit for multi-authored publications

Related to evaluation of researchers is the question of how to allocate credit for multi-authored publications. A general trend in academic publishing is the increasing number of authors per publication [Wuchty et al., 2007, Adams, 2012]. To illustrate this point, we have used the Microsoft Academic Graph (MAG, Chapter 4) to generate Figure 2.6. The figure

shows a change over time in the mean number of authors per publication across all publications found in the MAG which were published between the years 1900 and 2015.

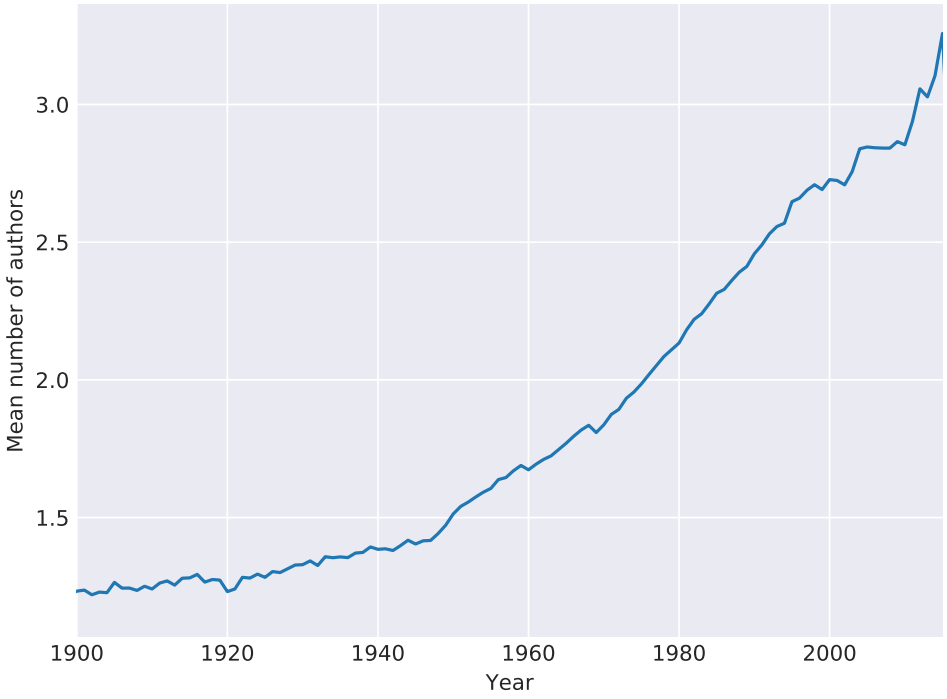

Figure 2.6: Mean number of authors per publication, averaged over a year, for publications from the MAG published between 1900 and 2015.

As science is becoming increasingly collaborative, there is a question of how to allocate credit to authors. Egghe and Rousseau [1990] and Van Hooydonk [1997] have discussed three methods. The simplest approach is to give each author full credit. In citation counting, this translates to each author receiving a full citation for each citation of every publication he or she authored. This is a common approach which is used for example by Google Scholar in their researcher profile page. The second possibility is allocating each author an equal fraction of the publication's citations. This approach has been advocated by Price [1981]. The third method is based on giving only the first author credit. Van Hooy-

donk [1997] has also proposed proportional counting, where each author is allocated a fraction of the credit based on their rank in the author list, with the first author receiving the most credit and the last author the least (with the basic fractional counting applied in case the authors are listed in alphabetical order). Van Hooydonk [1997] has compared the four approaches, and Waltman [2016] has provided a review of the recent work on different counting methods.

Evaluation of journals

In recent years there have been many attempts at creating new, more robust metrics that would complement or replace the JIF. Two of the most prominent metrics in this area are the *Eigenfactor*[2] [Bergstrom, 2007], created by the University of Washington, and the *SCImago Journal Rank*[3] (SJR) [González-Pereira et al., 2010], created by the SCImago lab.

The Eigenfactor algorithm works similarly as Google's search algorithm PageRank [Brin and Page, 1998]: the citations pointing to a journal are counted and weighted based on the ranking of the source journal. The source journal ranking is further normalised by the total number of citations that appear in that journal [Bergstrom, 2007]. The Eigenfactor metric thus helps to overcome one limitation of the JIF – the fact that JIF treats all citations as equal.

The SJR metric is similar to the Eigenfactor in that the SJR also weights the incoming citations based on the rank of the source journal so that citations from prestigious journals contribute more to the final rank than citations from less significant journals. The difference between the two metrics are the underlying data – the Eigenfactor uses the Thomson Reuters Web of Science database [University of Washington, 2017]

[2] `http://www.eigenfactor.org/`

[3] `http://www.scimagojr.com/journalrank.php`

(which is based on the Science Citation Index originally created by Eugene Garfield) while the SJR uses Elsevier's Scopus database [Scimago Lab, 2017].

A number of studies have provided a comparison of different journal evaluation metrics, including [Rousseau et al., 2009], [Franceschet, 2010], and [Kianifar et al., 2014] to name a few. A common finding among these studies is that these metrics tend to correlate quite well, but the correlations are not perfect; utilising these metrics as complementary information might therefore be useful. However, similarly as in the case of the h-index, JIF remains the most popular journal metric, with many journals reporting their JIF on their website.

Other units of evaluation

When it comes to other units of evaluation such as universities and countries, no methods similar to h-index and JIF focused specifically on these units exist. However, a number of public rankings have been produced, such as the SCImago Country Rank[4] and the Webometric Ranking of World Universities (Section 2.2.3). Furthermore, a number of studies focused on these unit exist [Csajbók et al., 2007, Lazaridis, 2010, Fakhree and Jouyban, 2011, Hassan and Haddawy, 2013].

Field normalisation of indicators

One concern about citation-based evaluation measures is the fact that citation patterns differ significantly across fields. For example, biochemical papers often contain many more references than mathematical papers which in turn leads to biochemical papers having higher average citation counts than mathematical papers [Moed, 2011]. This makes comparisons of outputs from different disciplines significantly harder or even

[4] `http://www.scimagojr.com/countryrank.php`

impossible. To enable such comparisons, various normalisation methods have been developed which aim to decrease or even eliminate the differences between fields. A similar situation happens when comparing publications published in different years as older publications had more time to attract citations than newer publications. Age normalisations are therefore also typically used.

Li et al. [2013b] divide the normalisation approaches into two main categories: *target-based* approaches which are functions of the cited papers and *source-based* approaches which are functions of the citing papers. The difference between these two classes is whether the weights or normalisation factors are functions of the cited papers (target-based) or of the citing papers (source-based). An example target-based normalisation method is normalising the citation count of each paper in a field s by the average number of citations received by papers in the field s [Li et al., 2013b]. The resulting value represents relative impact of a paper within its field. The calculation of the average value may or may not include un-cited publications [Li et al., 2013b]. Another possibility is normalising by median citation value [Li et al., 2013b]. An example of source-based normalisation is normalising the citation count of each paper in a field s by the average number of cited references per paper (average number of references found in each paper) in the field s. This approach is used in the *source normalised impact per paper* (SNIP) indicator [Moed, 2010], which is a metric for evaluating journal impact.

Waltman [2016] presented a different categorisation, and divided the normalisation approaches into two groups according on whether they are based on *average citation counts* (which were explained in the previous paragraph), or on *highly cited publications* (approaches where the proportion or the number of highly cited publications is used as the frame of reference). Several reviews and comparisons of the existing normalisa-

tion approaches exist, including [Waltman and van Eck, 2013], [Li et al., 2013b], and [Waltman, 2016]. Generally, there is no consensus on which of the normalisation methods is the most appropriate, and the choice of the method will therefore depend on the specific problem to be addressed.

New publication databases

Since the founding of the SCI, many new publication databases and citation indices have been created. Apart from SCI, which can be accessed through the Web of Science[5] (WoS), the other big commercial database is Scopus[6], which is owned and run by the largest academic publisher, Elsevier. Both WoS and Scopus offer APIs for accessing their data; however, both are commercial, and are available only to subscribers.

Possibly the largest index of scholarly publications and citations is Google Scholar[7]. Google Scholar provides a free search interface, but it does not offer an API and forbids crawling its search engine. As a result, research publication analyses that want to utilise Google Scholar data have to be done manually.

Because the underlying data used by these services differ (due to different focus, different collection mechanisms, etc.), the results and indicators provided by these services also differ. A number of studies have analysed these databases [Harzing, 2013, Franceschini et al., 2016, Khabsa and Giles, 2014] and provided comparisons of their data [Bar-Ilan, 2008, Falagas et al., 2008, Harzing and Alakangas, 2016]. A detailed review of the literature studying these databases is available in [Waltman, 2016]. In Chapter 4 we review a number of free and open alternatives to these three services.

[5] `https://webofknowledge.com`

[6] `https://www.scopus.com/`

[7] `https://scholar.google.com/`

Citation prediction

One type of studies which are relevant to evaluation of research publications are studies focusing on predicting future citation counts of publications. The goal of citation prediction is to use information about a publication to build a machine learning model for predicting citation counts the publication will receive in the future. While such models may not be directly applicable in research evaluation, they may be used to provide information about the importance of certain features for receiving high number of citations. One such study has focused specifically on identifying important features which influence future citation rates [Wang et al., 2011]. A similar study has been performed by Onodera and Yoshikane [2015] and Yan et al. [2012]. A number of different features have also been compared by Chakraborty et al. [2014] and Dong et al. [2015]. Furthermore, citation prediction was one of the tasks in the 2003 KDD Cup [Gehrke et al., 2003].

2.2.3 Web-based methods

In this section we focus on two sub-fields which make use of data from the Web, webometrics and altmetrics.

Webometrics

Webometrics is a relatively new research area, which has first been formally described in 1997 as the use of informetric and bibliometric approaches with online data in order to map the structure and usage patterns of the web [Almind and Ingwersen, 1997]. The underlying idea behind webometrics is that it is possible to replace papers and citations in the traditional citation networks with web pages and links between them [Almind and Ingwersen, 1997] (because of their similarity to cita-

tions, the links between web pages have sometimes been called "sitations" [Rousseau, 2003]). This analogy enables webometrics to use existing bibliometric and informetric methods, such as analyses of co-citation and bibliographic coupling (Section 2.2.1). In addition, utilising data from the Web enables tracking online scholarly communication, which offers new ways of assessing how research results are used by scientists, in teaching, and by the public [Björneborn and Ingwersen, 2004].

A prominent application of webometrics is the Webometric Ranking of World Universities[8] [Aguillo et al., 2008]. Björneborn and Ingwersen [2004] list four main areas of webometric research. Of these four, one (web technology analysis), which we do not mention here, is concerned mainly with studying the underlying technology rather than with the applications of webometrics in research evaluation. The three remaining areas of webometric research (the naming of the areas is from [Björneborn and Ingwersen, 2004], the explanations and examples are ours) are:

Analysis of the content of Web pages. An example topic belonging to this area is co-word analysis applied to Web pages. This approach has been used by Leydesdorff and Curran [2003] to identify the online connections between industry, universities, and government.

Analysis of the link structure. For example, a simple idea based on the link structure of the web is to evaluate the importance of a web page based on the number of links pointing to that site. This idea has been used to design a metric called the web impact factor (WIF) [Ingwersen, 1998] or to compare health web pages [Cui, 1999]. The link structure might also be useful for science mapping purposes [Harries et al., 2004].

[8] `http://www.webometrics.info/`

Web usage analysis. This area includes analysing log files of users' online behaviour. A significant correlation has been found between download counts of research articles and later citation impact [Brody et al., 2006].

Thelwall [2007] has provided a detailed review of the main research areas and developments in webometrics. A more recent review is available also in [Thelwall and Kousha, 2015a]. Another study has compared 39 scientific impact measures for evaluating journals (based on both citation and online usage data) in order to evaluate how they relate to each other and how well they represent scientific impact [Bollen et al., 2009].

One limitation of this approach is the fact that some research areas might be by nature more online-based then others (such as those where production of web pages and services is part of research) [Thelwall, 2007]. Metrics which utilise data from the web are also particularly susceptible to gaming. Priem and Hemminger [2010] point out that such attempts have occurred in the past with the goal of improving search engine results, though these have been successfully controlled (although not completely removed). One theoretical problem is the timely collection of data, as the Internet is constantly changing and growing. The collection of data from web search engines also poses several problems such as coverage issues or the question of how to ensure that all potentially relevant data have been retrieved. Finally, Priem et al. [2010] suggested that the Matthew effect of accumulated advantage could be at work on the Web.

Altmetrics

Altmetrics is the newest research area which was introduced in 2010 [Priem et al., 2010]; however, different altmetrics were investigated before the term was proposed. For example, Taraborelli [2008] has investigated

how data from social bookmarking services and online reference managers could be used for assessing semantic relevance and popularity of publications. The motivation for proposing altmetrics was the increasing difficulty in identifying relevant work among the growing amount of research and the limitations of the existing metrics, which often fail in this task [Priem et al., 2010]. Priem et al. [2010] have proposed altmetrics as a fast (compared to peer review and citation based metrics) alternative, and as a complementary method providing a broader view than the existing metrics.

Altmetrics is based on the idea of utilising data from the Web, particularly from social networks [Priem et al., 2010]. Researchers are increasingly discussing, linking, and bookmarking their work on various social networks, which brings an opportunity in the form of new data (such as Twitter mentions, online bookmarks, and blog posts) for measuring the impact of research. The difference between altmetrics and webometrics is in the underlying data used by these two fields – while webometrics mainly utilise the link structure and content of web pages, altmetrics focus on social media such as Twitter and Facebook. Webometrics therefore need to collect data through web crawling and web scraping or by utilising existing web indices and search engines [Almind and Ingwersen, 1997], whereas altmetrics typically work with Application Programming Interfaces (APIs) provided by the different social media services [Thelwall and Kousha, 2015b]. This is both an advantage and a disadvantage of altmetrics, because utilising APIs is a faster and a somewhat simpler method; however, this makes the data collection limited to what the APIs offer [Bornmann, 2014] and creates a need for a separate program for each API. A recent review by Erdt et al. [2016] has identified two major research directions in altmetrics: *cross-metric validation* and *coverage of altmetrics*. Cross-metric validation studies focus on comparing altmetrics

with other metrics, especially citation counts. A positive correlation with citation counts is considered to be evidence of the value of an indicator [Thelwall and Kousha, 2015a]. For example, Costas et al. [2015] have conducted a cross-disciplinary comparison of different altmetrics with citations. They found the correlations to be positive but relatively weak. A similar study was conducted by Thelwall et al. [2013]. Li and Thelwall [2012] have compared F1000 ratings and Mendeley reader counts with citation counts. They found significant correlations between both metrics and citation counts, with the correlations for Mendeley reader counts much stronger than for F1000 ratings.

The second group of studies identified by Erdt et al. [2016] (studies focused on coverage of altmetrics) investigate the number of research articles for which different altmetrics are available. Most studies have generally found the coverage of altmetrics to be low, with the highest coverage offered by Mendeley (59.2% across 15 studies investigated by [Erdt et al., 2016]) and by Twitter (24.3% across 11 studies) [Erdt et al., 2016].

Another recent review of altmetrics is available in [Thelwall and Kousha, 2015b]. Priem and Hemminger [2010] have summarised existing databases and services which can be used for collecting altmetric data. These sources include social bookmarking services, reference managers, blogs, microblogs, and comments on articles. [Bornmann, 2014] has provided a review focused on summarising the main advantages and limitation of altmetrics. Due to the reliance on data from the Web, altmetrics share some limitations, such as susceptibility to gaming and data collection issues, with webometrics.

2.2.4 Text-based methods

In this section we review the existing works in bibliometrics and related areas which make use of text for the evaluation of research publications and for other relevant tasks. The recent growth of Open Access publishing has created a new opportunity in this area, which has already led to the creation of a number of open datasets of research publications available online (Chapter 4).

Open Access (OA) is the practice of providing free unrestricted access to scholarly literature. In contrast to the traditional subscription based journal literature, OA removes fees for accessing the literature as well as most copyright and licensing restrictions. OA was defined in three public statements, the 2002 Budapest Open Access Initiative (BOAI) [Chan et al., 2002], the 2003 Bethesda Statement on Open Access Publishing [Brown et al., 2003], and the 2003 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities [Bullinger et al., 2003]. An important part of the OA movement is that it enables harvesting the full text of the articles and processing them automatically using computer software. This means the article full texts can be used in bibliometric analyses. Furthermore, the UK government has recently accepted a policy which states that since 2014, all publicly funded research has to be published as OA [Research Councils UK, 2012] using either gold (publishing in OA journals) or green (self-archiving in author's institutional repository) OA options. In 2013, the USA has announced a similar policy [Office of Science and Technology Policy, 2013].

The proportion of OA literature has been found to be around 20% in 2009 [Björk et al., 2010], while a report from 2013 states that the proportion of OA articles from 2011 is almost 50% [van Noorden, 2013]. The OA movement has already manifested itself in bibliometrics, as it has been shown OA publications tend to have higher citation counts,

page views, and generally attract more attention [Harnad and Brody, 2004, Eysenbach, 2006, Wang et al., 2015]. However, in our view, the most significant change will happen thanks to opening of the data to the public.

In the remainder of this section we review the approaches to automated research evaluation which in some way utilise text. In particular, we focus on approaches which utilise co-word analysis to map scientific disciplines, approaches which focus on analysis and classification of citation contexts, and on other approaches, such as applications of text-based clustering in citation normalisation. An important distinction between the different text-based approaches is whether they have utilised titles, abstracts, or full text content of publications. While some methods, such as the methods introduced in the following section which analyse co-occurrence of words in text (co-words) may work with as little as just titles, some other methods (particularly methods which analyse citation contexts) require access to full text content of scientific documents. This is an important distinction which may affect whether a method can be applied in a certain context. Where relevant, we therefore note whether a method or a set of methods utilise titles, abstracts or full text.

Co-word analysis and science mapping

Bibliometrics is not restricted to using only citations. Possibly the best known bibliometric technique based on text is co-word analysis, a technique similar to the bibliographic coupling and co-citation analysis, which was introduced by Callon et al. [1983]. Co-word analysis was proposed as a method for capturing the strength of the relation between documents and is based on the extraction of co-words from scientific articles – pairs of words which both appear in the same document. The co-words can be extracted from any part of the document – the title, abstract, keywords,

or full text. The extracted co-words are then analysed to identify hierarchies or clusters of words that appear frequently together. Co-word analysis is frequently used in *science mapping*, a field of study concerned with identifying and mapping the relations between scientific disciplines as well as tracking and visualising the evolution of disciplines [Börner et al., 2003].

Using words instead of citations has several advantages. Words are meaningful and ubiquitous, and unlike citations, they do not take time to accumulate [Leydesdorff, 1989]. As [De Bellis, 2009] has pointed out, using words also requires less assumptions than using citations (such as assumptions about the reasons to cite an article). On the other hand, co-word analysis has been criticised because of the varying quality of used keywords and index words [He, 1999] and because single words used in the analysis lack the meaning of the context [Leydesdorff, 1989].

Different indices for measuring the strength of relationship between words have been used for the clustering. One of the earliest and simplest indices is the *inclusion index* [De Bellis, 2009], which is defined as [Callon et al., 1983]:

$$I_{ij} = \frac{c_{ij}}{c_i}, \tag{2.2}$$

where $c_i < c_j$, c_{ij} is the co-occurrence of words i and j, c_i is the occurrence of the word i, and c_j is the occurrence of the word j. The inclusion index captures the probability of finding the word i in an article given that the word j is already present [De Bellis, 2009]. If $I_{ij} = 1$, i is present in every article in which j is present, which in co-word analysis is interpreted as full inclusion of the first word by the second (hence the name) [De Bellis, 2009]. The inclusion index can therefore detect hierarchies of topics in a field [Wang et al., 2012]. Callon et al. [1983] have used the inclusion index to create a map of topics within a medical

science field.

Callon et al. [1983] observed the pattern of inclusion is not a typical case, such as when new and developing fields are concerned. Some words may have occurred very infrequently but have a significant relationship with other words. To capture this pattern, Callon et al. [1983] have proposed the *proximity index*:

$$P_{ij} = \frac{c_{ij}}{c_i c_j} \cdot N, \tag{2.3}$$

where c_{ij}, c_i, and c_j are defined as in the inclusion index (Equation 2.2), and N is the number of documents in the collection. The proximity index captures the word pair frequencies that point to minor (and potentially growing) topics [Rip and Courtial, 1984].

To capture the pattern of mutual inclusion of words, the *equivalence index* has been defined [Turner et al., 1988, Callon et al., 1991]:

$$E_{ij} = \frac{c_{ij}}{c_i} \cdot \frac{c_{ij}}{c_j} = \frac{c_{ij^2}}{c_i c_j}, \tag{2.4}$$

where c_{ij}, c_i, and c_j are defined as in Equations 2.2 and 2.3. The equivalence index E_{ij} has a value between 0 and 1, and, similarly to the inclusion index, measures the probability of a word i appearing in a document given that j is already present, and, inversely, the probability of a word j appearing in a document given that i is already present [He, 1999].

Other measures of similarity between words which have been used in co-word analysis include the *Jaccard index*, $J_{ij} = \frac{c_i j}{c_i + c_j - c_{ij}}$ [Rip and Courtial, 1984] and the *cosine similarity* [Salton and McGill, 1986], represented in co-word analysis as $S_{ij} = \frac{c_{ij}}{\sqrt{c_i \cdot c_j}}$ [Peters and Van Raan, 1991, Peters and van Raan, 1993], which is a different form of the equivalence index (Equation 2.4). A comparison of different similarity coefficients for co-word analysis has been presented by Sternitzke and Bergmann [2009].

To use co-word analysis in science mapping, techniques such as cluster analysis, community detection, and dimensionality reduction are used to identify important words, word hierarchies, and clusters [Cobo et al., 2011]. These methods include *principal component analysis* (PCA) [Wold et al., 1987], *multi-dimensional scaling* (MDS) [Kruskal, 1964], and various *clustering algorithms*, applied both to co-words (and the coefficients described above) and to document vectors produced using methods such as *vector space models* [Salton et al., 1975] and *latent semantic analysis* (LSA) [Deerwester et al., 1990]. A detailed review of methods used in science mapping including the ones mentioned above is provided in [Börner et al., 2003]. Cobo et al. [2011] have also summarised existing science mapping tools.

Leydesdorff and Hellsten [2005] have used co-word analysis to analyse the publications, patents, and newspaper articles on stem cell research. Co-word analysis has also been applied to MEDLINE[9] keywords to analyse complementary but disjointed literature [Stegmann and Grohmann, 2003], to characterise relations between science and technology [Noyons and van Raan, 1994, Bhattacharya et al., 2003], and to map scientific fields [Braam et al., 1991, Peters and van Raan, 1993, Ding et al., 2001, Lee and Jeong, 2008] and relations between fields [Onyancha and Ocholla, 2005].

Analysis and classification of citation contexts

One area of computational linguistics and natural language processing that is related to the evaluation of research publications is the area concerned with the analysis and classification of citation contexts. In bibliometrics and related areas, the use of citations for impact analysis is usu-

[9]MEDLINE is a database of biomedical literature accessible through PubMed, a free online search engine.

ally based on the assumption that all citations are equal, and a citation from publication a to publication b is interpreted as influence of publication b on publication a. However, it has been shown acknowledging the influence of prior work is only one of many reasons for citing a publication [Nicolaisen, 2007, Bornmann and Daniel, 2008]. A typical goal of works focusing on citation contexts is distinguishing citations mentioned in different contexts and analysing and identifying the different reasons and motives for citing. This set of methods therefore by definition requires the access to full text content of publications.

The details of where and how frequently citations appear in text has been of interest to a number of researchers [Hou et al., 2011, Bertin et al., 2013, Bertin and Atanassova, 2014, Bertin et al., 2016b,a, Bertin and Atanassova, 2016, Ding et al., 2013, Hu et al., 2015, Atanassova and Bertin, 2016]. The focus is typically on examining whether differences in the use of references in text can be identified and whether these differences can help in assessing the value of the reference. [Hou et al., 2011] have analysed how frequently similar (defined as having 10 or more references in common with the citing paper) and dissimilar (with less than 10 references in common with the citing paper) papers are repeatedly referenced in text. They found that similar publications tend to appear repeatedly in the text, and that the difference between the recurrence of similar and dissimilar publications (which appear less often) is statistically significant. Based on their observations, the authors have suggested counting citations in text may be a more accurate measure of scientific contribution. Interestingly, they also found their citation counting method decreases the rank of journals with a high proportion of review articles. Ding et al. [2013] have studied the distribution and recurrence of references in scientific articles with respect to the different sections found in scientific articles. Their specific focus was on analysing whether

counting all occurrences of a reference in a text produces different rankings from counting each reference once. They found that for highly cited references both methods produce similar ranks, but for the remainder of the references they differ significantly. A similar study has been done by [Hu et al., 2015] and by Bertin et al. [2013], who have compared the distribution of references in scientific articles with the IMRAD (Introduction, Method, Results, and Discussion) structure, a format typically followed by scientific articles. In follow-up studies, the authors have analysed the use of verbs [Bertin and Atanassova, 2014] and n-grams [Bertin et al., 2016b] found in citation contexts across the four sections of the IMRAD structure as well as the age of the references found across the four sections [Bertin et al., 2016a]. They have also identified negation (such as disagreeing with previous findings) in scientific publications and analysed its relation to scientific citations [Bertin and Atanassova, 2016]. They found that negational contexts most frequently appear in the discussion section, and that a significant portion of negational contexts do not occur together with a reference. Atanassova and Bertin [2016] have also provided an in-depth analysis of recurring references.

A number of researchers have also explored the possibilities around automated classification of the function and sentiment of citations [Abu-Jbara et al., 2013, Agarwal et al., 2010, Athar and Teufel, 2012a,b, Butt et al., 2015, Di Iorio et al., 2013, Jurgens et al., 2016, Lauscher et al., 2017, Li et al., 2013a, Liu et al., 2015, Pride and Knoth, 2017, Teufel et al., 2006, Valenzuela et al., 2015, Wan and Liu, 2014, Xu et al., 2015, Zhang et al., 2013, Zhu et al., 2015]. The underlying idea is to use features extracted from the context of each reference found in the citing publication, and in some cases from other parts of the publication text or metadata, to create models for automatically classifying each outgoing citation according to its function within the citing publication. Pride

and Knoth [2017] have summarised the steps needed to train such a classifier. These steps include (1) text extraction, (2) parsing the full text to detect positions of references and other parts of the document, (3) feature extraction, (4) training and applying a classifier.

One of the first steps for most studies has been defining the classification scheme. Three types of classification schemes are typically used [Jurgens et al., 2016]: (1) schemes focused on the centrality of the citation (whether the referenced publication is necessary for understanding the citing publication, or whether it is used to position the work within a broader context; this classification scheme was used for example by Valenzuela et al. [2015], Jurgens et al. [2016], and Li et al. [2013a]), (2) schemes focused on citation function (particular purpose of the citation, such as to provide background, or to support a statement; this classification scheme was used for example by Teufel et al. [2006], Jurgens et al. [2016], and Agarwal et al. [2010]), and (3) schemes focused on citation sentiment (whether the citation is referenced in a positive, negative, or a neutral context; this classification scheme was used for example by Athar and Teufel [2012a], Abu-Jbara et al. [2013], and Lauscher et al. [2017]).

Several studies have looked at the problem of citation context identification. Abu-Jbara et al. [2013] have utilised Conditional Random Fields (CRF) for this task. Athar and Teufel [2012b] and Valenzuela et al. [2015] have also focused on the problem of detecting implicit references (in-text references which do not contain an explicit link to a publication but instead mention a method or an author name). Pride and Knoth [2017] have compared the output of several libraries for publication text parsing, with focus on the number of references detected by each library.

A wide variety of features have been used by different researchers. Pride and Knoth [2017] have divided the features used by different studies into two categories depending on whether they rely on *internal* (ex-

tracted from the full text of the publication) or *external* (extracted from additional, external information) information. The internal features include number of times a publication is referenced in the text of the citing publication [Valenzuela et al., 2015, Pride and Knoth, 2017], position of the reference (such as which section does it appear in) [Abu-Jbara et al., 2013, Jha et al., 2017], and various lexical and morphological features [Jurgens et al., 2016, Teufel et al., 2006]. The external features include sentiment lexicons [Butt et al., 2015], author information [Valenzuela et al., 2015, Jha et al., 2017], and similarity with the cited publication [Zhu et al., 2015, Valenzuela et al., 2015]. A number of different models have been used for training the classifier, with support vector machine (SVM) [Valenzuela et al., 2015, Abu-Jbara et al., 2013, Agarwal et al., 2010, Athar and Teufel, 2012a, Lauscher et al., 2017], Naïve Bayes (NB) [Abu-Jbara et al., 2013, Agarwal et al., 2010], and random forests (RF) [Jurgens et al., 2016, Valenzuela et al., 2015] being among the most popular models.

One specific challenge is inherent in this research problem. To be able to train automated models, a set of correct labels for training is needed. However, labelling a citation between two publications according to the function the citation plays in the citing paper requires an annotator familiar with the subject area who can identify and understand the citation contexts within the citing publication to identify the function of each outgoing citation. As a result, creating labels for citation classification is a time and resource intensive task. Most existing approaches are therefore limited to a single discipline. Abu-Jbara et al. [2013], Athar and Teufel [2012a], Butt et al. [2015], Jha et al. [2017], Jurgens et al. [2016], Pride and Knoth [2017], Teufel et al. [2006], Valenzuela et al. [2015], Wan and Liu [2014], Zhu et al. [2015] have used articles from the Natural Language Processing (NLP) domain, while Li et al. [2013a], Liu et al.

[2015], Xu et al. [2015] have worked with articles from the biomedical domain. Lauscher et al. [2017] have worked with labelled data from a single discipline (NLP), but trained their model using embeddings created using a multi-disciplinary dataset. Jurgens et al. [2016] have trained their classification model on a smaller set of labels and used the trained model to analyse citation patterns across a much larger dataset covering a large proportion of the NLP discipline. To the best of our knowledge, only Valenzuela et al. [2015], Jurgens et al. [2016], and Jha et al. [2017] have publicly released their annotated datasets; however, at the time of writing this section, the link to the dataset provided by Jurgens et al. [2016] did not work. Another two challenges of citation classification which were highlighted by Pride and Knoth [2017] are the difficulty of extracting some complex features from publication full texts and the errors produced by the existing libraries for converting PDF files to text (which may create error in the classification).

Citation contexts have also been utilised in other tasks. Ritchie [2009] made use of citation context to improve information retrieval, and Siddharthan and Teufel [2007] used citation contexts for attribution of expressions to scientific publications. The context of references found within citing papers has also been used for enhancing author co-citation analysis [Jeong et al., 2014], in paper summarisation [Abu-Jbara and Radev, 2011], in citation recommendation [Kataria et al., 2011], for topic classification [Caragea et al., 2015], and (in combination with PageRank) for publication ranking [Liu et al., 2013]. Zhang et al. [2013], Ding et al. [2014], and Radoulov [2008] have provided surveys of the area of automated classification of citations.

Other applications of text in scientometrics

Text analysis has also been used in other tasks. Yan et al. [2012] have used semantic distance between a publication and its references (which was calculated using abstracts) to predict future citations, and Whalen et al. [2015] have used semantic distance (calculated using full text) between a publication and the publications that cited it for the same task. Holste et al. [2011] and Hörlesberger et al. [2013] have used publication full text to identify "frontier research" in research project proposals. Kostoff et al. [2001] have use text clustering (based on abstracts) to identify topical communities among citing publications to characterise the communities which have referenced a publication. Glenisson et al. [2005] have combined text- and citation-mining to create a map of a scientific area and compared results obtained using abstracts and full text. The found abstracts and full text produce somewhat different clustering results. For clustering using full text, they suggest parsing the publications to remove sections which are not relevant. Colliander [2015] has combined bibliographic coupling and content similarity to improve citation normalisation, and he has shown the content-based approach (which has utilised publication abstracts) outperforms other methods. Wang et al. [2012] and Feng et al. [2017] have integrated semantic relationships into co-word analysis. Gerrish and Blei [2010] have use a dynamic topic model (DTM) to measure thematic changes in a collection over time. This was used to measure the importance of individual documents (their influence on the topics discussed in the other documents) within the collection. This is an interesting approach which does not require citation information. However, in this case, for a number of documents discussing the same idea, it might be difficult to recognise which document or documents were the main influencers of the field. A similar task was explored by Glänzel et al. [2017], who studied the changes in the vocabulary of a

selected set of publications over the period of three decades. Livne et al. [2013] and McKeown et al. [2016] have used features extracted from publication full text for citation prediction. Text mining has also been used to analyse funding patters over time [Park et al., 2016] using abstracts of project proposals.

2.3 Research evaluation initiatives

In Chapter 1 we have introduced a number of scenarios demonstrating the growing interest in the evaluation of research outcomes. Due to this growing interest and the increasing availability of data pertaining to research (especially research publication), the use of various research metrics is becoming widespread. However, due to the limitations of the existing metrics, there are concerns about the applications of these metrics. This is because it is not uncommon for research metrics to be used in scenarios for which they were not designed. For example, thanks to the free online academic search engine, Google Scholar, it has become very easy to obtain a researcher's h-index value. As a consequence, the h-index is being reported on scientists' résumés [Ball, 2007] and used by hiring committees [Acuna et al., 2012] despite scientists urging caution when using the index for this purpose [Kreiner, 2016].

A number of initiatives and reviews have recently emerged which discuss these issues and provide suggestions for a better use of research metrics. One of the first has been the San Francisco Declaration on Research Assessment (DORA) [San Francisco DORA, 2012]. The creation of DORA has been motivated by the increasing use of the JIF for evaluation of individual articles published in a journal, and the issues associated with this practice [San Francisco DORA, 2012]. DORA provided a number of suggestions for improving research assessment, such as to

stop using journal-based metrics in funding, appointment, and promotion considerations; to assess research based on its own merits (rather than on the basis of the venue where it was published); and to make a better use of the opportunities provided by publishing research articles online.

The Leiden Manifesto [Hicks et al., 2015] was created with a similar aim in mind. The increasing misuse of research metrics has motivated the authors to provide recommendations and best practices for metrics-based research assessment. Their recommendations include keeping data collection and analysis open and transparent, which will enable those being evaluated to verify the data and analysis; measuring performance against the research mission of the institution, group, or individual, so that the choice of indicators is based on context; and using quantitative evaluation to support expert assessment to challenge bias [Hicks et al., 2015].

The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management [Wilsdon et al., 2015] was created to provide an independent review of the use of metrics in research assessment. It focused on all aspects of the use metrics in research evaluation including benefits and limitations, effects on research culture, gaming of metrics, and the use of metrics in the UK Research Excellence Framework. The report has proposed the notion of *responsible metrics*, a framework for the development and use of research metrics which encourages appropriate uses of metrics. Responsible metrics have been described by the report in terms of five dimensions [Wilsdon et al., 2015]:

- "**Robustness:** basing metrics on the best possible data in terms of accuracy and scope,

- **Humility:** recognising that quantitative evaluation should support – but not supplant – qualitative, expert assessment;
- **Transparency:** keeping data collection and analytical processes open and transparent, so that those being evaluated can test and verify the results;
- **Diversity:** accounting for variation by field, and using a range of indicators to reflect and support a plurality of research and researcher career paths across the system;
- **Reflexivity:** recognising and anticipating the systemic and potential effects of indicators, and updating them in response."

The report has also provided a number of recommendations. Many of the recommendations, including asking for transparency and openness of data and methods, and reducing emphasis on journal impact factors, were similar to the recommendation provided by [Hicks et al., 2015]. The report has, among other things, called for a better use of existing data and information sources, and increased funding in research information infrastructure.

The Science Europe Position Statement [Science Europe, 2016] is the most recent of the initiatives and reviews mentioned in this section. The statement was predominantly focused on data. The motivation behind creating the statement was increasing the interoperability of research information systems. The report has provided several recommendations and guidelines in that area, which were summarised into four core principles [Science Europe, 2016]: (1) flexibility (systems should allow extensions in terms of data, external sources, etc.), (2) openness (data should be available for external use), (3) FAIRness (foster findability, accessibility, interoperability, and reusability), (4) minimising data entry (avoid having to enter data multiple times).

It can be seen a number of recommendations appear across multiple reports. These are particularly the recommendations concerned with reducing the use of journal-based metrics, evaluating with respect to context, and improving openness of data and methods. The recommendations provided by these reports influence how we think about research evaluation and how we implement our methods.

2.4 Summary and discussion

Evaluation of research publications is becoming increasingly more important. In this section we have reviewed the existing approaches to evaluation of research publications and highlighted their strengths and weaknesses. These approaches can be broadly categorised as (1) citation-based (bibliometric) approaches, (2) web-based (webometric and altmetric) approaches, and (3) text-based approaches.

The citation-based approaches rely on citation data pertaining to publications, authors, journals, or other evaluation units. A large number of indicators utilising this data have been proposed, including the journal impact factor and the h-index. The citation-based methods are typically based on the assumption that a citation between two publications represents the influence of the cited work on the citing work. However, it has been shown there are many reasons why one might want to cite a publication [Nicolaisen, 2007, Bornmann and Daniel, 2008]. In fact, most citations are not essential for the citing publication and in many cases can be substituted [Simkin and Roychowdhury, 2002, Valenzuela et al., 2015, Ricker, 2017]. Furthermore, citations are often used as a proxy to research quality [Bornmann and Haunschild, 2017]; however, the relation between citations and research quality, which has been extensively studied [Aksnes, 2003, Antonakis et al., 2014, Bornmann and Leydesdorff,

2015], is unclear.

The web-based approaches, on the other hand, rely on data extracted from the Web, such as web pages and links between them and data from social media and other online services. The advantage of these methods is that they provide a different view of the uptake of research. Nevertheless, because the web-based approaches are typically based on counting the interactions in the scholarly communication network, they are affected by similar limitations as the citation-based approaches. For example, there is a lack of evidence demonstrating what these approaches capture and whether and how they can be linked to publication quality [Bornmann, 2014]. Furthermore, there are a number of aspects concerning data collection and quality [Priem et al., 2010, Bornmann, 2014].

In addition to the aforementioned limitations, the citation-based and web-based approaches are fully dependent on *external evidence of publication usage*. However, as has been shown, assessing the value of a piece of work solely on the number of interactions often does not provide sufficient evidence of quality and value.

The above limitations have led to the emergence of various methods that utilise publication content. Two main directions to utilising text in evaluation of research publications have been described in this chapter – co-word analysis and citation context analysis. Co-word analysis replaces citation and web links with words extracted from scientific documents. Within this area, the focus has been predominantly on analysing the relations between scientific documents and on utilising these relations to map scientific fields and their relations.

Citation context analysis aims at overcoming one of the largest limitations of the citation-based approaches – the fact that not all citations express the influence of the cited paper on the citing paper. Citation context analysis has been used to create automated methods for cita-

tion classification. This approach however relies on the ability to access the full text content of the citing publications and the ability to extract citation contexts and complex features from these publications. Furthermore, these approaches need to be *trained*, and as a consequence, require labelled data. However, collection of labels for training may be challenging, especially if multiple disciplines are concerned. Aside from the automated citation classification approaches, there is a lack of methods utilising text directly applicable to evaluation of research publications.

Overall, based on our review in this chapter, a serious limitation of the citation- and web-based approaches is that they rely on external evidence. Furthermore, while a number of researchers have successfully made use of text for various related tasks, significantly fewer studies have focused specifically on developing new methods which would utilise text to provide more robust and reliable metrics, and the existing studies applicable in this area have been largely limited to studying and classifying citation context. However, we believe text analysis offers many more opportunities for improving the existing metrics and developing new metrics. Together with the limitations of the citation- and web-based methods, this lack of existing text-based methods constitutes the motivation behind the research work presented in this thesis and forms the main research question investigated in this thesis:

> **How to effectively incorporate publication content into research evaluation to provide additional evidence of publication quality?**

In the following chapters we introduce new approaches and solutions which address the the above limitations and experiments evaluating the performance of our methods.

Part II

Evaluation of Research Publication

Chapter 3

The concept of research publication quality

> *Not everything that counts can be counted, and not everything that can be counted counts.*
>
> – Albert Einstein

"Quality" is a commonly used term in research evaluation. It has been stated the goal of peer review is ensuring only high-quality research gets published [Kelly et al., 2014], and the focus of evaluative scientometrics is on measuring the quality of published research [Bornmann and Haunschild, 2017]. However, what exactly is research quality? In scientometrics, quality has typically been measured in terms of the number of citations [Butler, 2008, Abramo et al., 2010, Bornmann and Haunschild, 2017], nevertheless, many researchers have pointed out issues associated with making such a connection [Meho, 2007, Adler and Harzing, 2009, Adler et al., 2009, MacRoberts and MacRoberts, 2010, Onodera and Yoshikane, 2015, Ricker, 2017]. The reasons why the connection between citation counts and quality are considered problematic are many, from the fact citations may be used to criticise as well as praise [Onodera

and Yoshikane, 2015] to the fact quality is a complex and multi-faceted concept which cannot easily be expressed in a single indicator [Ricker, 2017]. Peer review, especially when it comes to journals with high impact factor, is often considered to be the best available measure of quality [Garfield, 2003, Bornmann and Daniel, 2005, Kreiman and Maunsell, 2011]; however, this method of recognising high quality research also has its drawbacks, including reviewer bias [Teixeira da Silva and Dobránszki, 2015], and the fact reviewers often do not agree on which papers are the best and deserve to be accepted [Francois, 2015].

Nevertheless, if we wish to measure the quality of research outputs, the first thing we need to do before choosing specific metrics is to discover the dimensions of the concept. Once we have a better understanding of research quality, we can develop methods for assessing some of its dimensions. This chapter addresses this question, i.e.:

> RQ1: *What is research publication quality and what factors influence it?*

This chapter is dedicated to surveying existing definitions and analyses of the concept of research publication quality. We start by reviewing the criteria used in several national evaluation exercises (Section 3.1), and in journal peer review (Section 3.2). Next, we look at existing studies of criteria influencing research publication quality (Section 3.3). The rest of the chapter (Section 3.4) is devoted to presenting the results of a survey which we conducted at the Open University with the aim of gaining a better understanding of the perception of research quality among scientists. We summarise our finding and conclude the chapter in Section 3.5.

3.1 Research evaluation frameworks

Systematic research assessment has become an important aspect of output analysis and decision-making for many governments. As of 2010, at least 33 have some form of a university ranking system [Hazelkorn et al., 2010] and at least 14 countries have implemented some form of a performance-based research funding system [Hicks, 2012]. These exercises typically focus on reviewing the quality and impact of research done at publicly funded research institutes (mainly universities) across the country. The results of these assessment exercises are typically used to track performance of these institutes, provide evidence of value to taxpayers, and in case of the performance-based funding systems, determine funding allocation.

Several previous studies have provided comparative assessments of various national evaluation systems [Geuna and Martin, 2003, Hazelkorn et al., 2010, Hicks, 2012, Rijcke et al., 2015]. Geuna and Martin [2003] have examined evaluation systems used across 12 countries in Europe, Asia and Pacific, with particular focus placed on the United Kingdom. Interestingly, they reported that while in the short term, national research evaluation exercises can increase efficiency, in the long term, after a number of exercises have been performed, they may lead to diminishing returns due to high costs and decreasing benefits from repeated evaluation. Hazelkorn et al. [2010] has produced a report for the European Commission which has reviewed systems used for the evaluation of universities used around the world, including those operated by commercial organizations and the media. The focus of the report is on summarizing aims of the performed evaluations, indicators used, levels of evaluation, target users and other related information. Hicks [2012] has provided an in-depth analysis of 14 individual performance-based research funds,

focusing of studying common themes with the aim of identifying those which could inform innovation policy. The review by Rijcke et al. [2015] focuses largely on the use of quantitative metrics within different evaluation systems.

As none of the existing studies have analysed the criteria used to assess research quality, we briefly review several existing national evaluation systems with focus on quality criteria introduced by these systems. Namely, we review research evaluation systems of the following countries: the United Kingdom (Section 3.1.1), Australia (Section 3.1.2), New Zealand (Section 3.1.3), Italy (Section 3.1.4), and the Netherlands (Section 3.1.5). These five frameworks were selected for review according to the following criteria, which are similar to the criteria used by Hicks [2012] for selecting performance-based research funding systems for review:

- They evaluate research (rather than focusing on teaching or quality of degrees).
- They evaluate research outputs (rather than focusing purely on size of the institute or incoming funding).
- They focus on published outputs rather than research proposals (perform a review of past outcomes rather than funding proposals).
- They include a peer review component (as opposed to systems based purely on quantitative indicators).
- They are national evaluation systems (rather than evaluations performed by individual institutes, companies or media).
- They were performed more than once (they are not in a testing/partial implementation stage).

Our final constraint in reviewing the evaluation systems was language, as we were only able to review those described in English language. The aim of this review is not to be exhaustive in terms of inclusion of all known national evaluation systems, but rather to provide an overview of quality criteria used in some of the best known and better established research evaluation systems.

After a short description of each of the evaluation systems, we list the criteria used in evaluating research outputs submitted for the evaluation. While talking about each of the frameworks, we review the latest completed evaluation exercise. We don't review exercises scheduled for the future, as guidelines for those exercises may change. We also don't review the older versions of the exercises, as we consider the latest version to be the best developed one. At the end of the section, we provide a summary of the quality criteria and point out similarities and differences between the criteria used by different systems (Section 3.1.6).

3.1.1 United Kingdom

The UK's Research Excellence Framework (REF) is a system for assessing the quality of research done in UK higher education institutions (HEI). REF replaced the previous Research Assessment Exercise (RAE) [Research Excellence Framework, 2012], which was conducted several times since 1986. REF is managed by the Higher Education Funding Council for England (HEFCE), and was performed once in 2014, focusing on research outputs from 2008–2013. It was based primarily on peer review, however, the use of quantitative indicators (particularly citation counts) was permitted as a support for peer review judgements. In addition to research outputs (up to four per each member of staff included in the submission; accepted types of research outputs included journal and conference publications, books, design, software, data, and other types

of outputs), the panels also took into account other information, such as funding and details of awarded Ph.D. degrees.

For the evaluation, research disciplines were distributed across four broad panels (A-D). The evaluation of submitted research outputs constituted 65% of the overall score and was done according to the following criteria [Research Excellence Framework, 2012]: (1) *originality*, (2) *significance*, (3) *rigour*. However, the general assessment guidelines didn't provide a common definition or a description of the criteria and each panel was asked to provide their own interpretation of the criteria. Panel A, which covered medicine, health and life sciences, specified the following characteristics of quality, at least one of which was required to meet the definition of research used for REF [Research Excellence Framework, 2012]:

- "scientific rigour and excellence, with regard to design, method, execution and analysis
- significant addition to knowledge and to the conceptual framework of the field
- potential and actual significance of the research
- the scale, challenge and logistical difficulty posed by the research
- the logical coherence of argument
- contribution to theory-building
- significance of work to advance knowledge, skills, understanding and scholarship in theory, practice, education, management and/or policy
- applicability and significance to the relevant service users and research users

- potential applicability for policy in, for example health, healthcare, public health, animal health or welfare".

Panel B, which covered physical sciences, engineering and mathematics, defined originality, significance and rigour as follows [Research Excellence Framework, 2012]:

- "**Originality** will be understood as the extent to which the output introduces a new way of thinking about a subject, or is distinctive or transformative compared with previous work in an academic field.

- **Significance** will be understood as the extent to which the work has exerted, or is likely to exert, an influence on an academic field or practical applications.

- **Rigour** will be understood as the extent to which the purpose of the work is clearly articulated, an appropriate methodology for the research area has been adopted, and compelling evidence presented to show that the purpose has been achieved."

Panel C, which covered social science disciplines, provided the following interpretation of the generic criteria for assessing outputs:

- "**Originality** will be understood in terms of the innovative character of the research output. Research outputs that demonstrate originality may: engage with new and/or complex problems; develop innovative research methods, methodologies and analytical techniques; provide new empirical material; and/or advance theory or the analysis of doctrine, policy or practice.

- **Significance** will be understood in terms of the development of the intellectual agenda of the field and may be theoretical, methodological and/or substantive. Due weight will be given to potential

as well as actual significance, especially where the output is very recent.

- **Rigour** will be understood in terms of the intellectual precision, robustness and appropriateness of the concepts, analyses, theories and methodologies deployed within a research output. Account will be taken of such qualities as the integrity, coherence and consistency of arguments and analysis, such as the due consideration of ethical issues."

Finally, panel D, which included arts and humanities, provided the following definitions of the assessment criteria:

- "**Originality:** a creative/intellectual advance that makes an important and innovative contribution to understanding and knowledge. This may include substantive empirical findings, new arguments, interpretations or insights, imaginative scope, assembling of information in an innovative way, development of new theoretical frameworks and conceptual models, innovative methodologies and/or new forms of expression.

- **Significance:** the enhancement or deserved enhancement of knowledge, thinking, understanding and/or practice.

- **Rigour:** intellectual coherence, methodological precision and analytical power; accuracy and depth of scholarship; awareness of and appropriate engagement with other relevant work."

It can be seen that the definitions provided by the panels in some cases significantly differ. For example, panel C is the only panel that mentions consideration of ethical issues as part of their interpretation of rigour. On the other hand, panel D is the only panel that included "awareness of and

appropriate engagement with other relevant work" in their interpretation or rigour. As the definitions differ across panels, it is unclear whether certain criteria, such as the use of references mentioned by panel D, play a significant role also within the other panels. However, in our view, it seems most likely to expect the parts of the interpretations which are shared across all panels to be an essential aspect of the evaluation for all panels, while the parts that are unique to each panel play less of a central role.

3.1.2 Australia

The Australian national evaluation exercise, Excellence in Research for Australia (ERA), is managed by the Australian Research Council (ARC) [Australian Research Council, 2015b]. ERA replaced the previous Research Quality Framework and was so far performed in 2010, 2012 and 2015, with the next round scheduled for 2018 [Australian Research Council, 2017].

According to the evaluation handbook [Australian Research Council, 2015a], unlike the UK REF, which is primarily based on peer review judgements, ERA is based on the principle of expert review informed by quantitative indicators. For purposes of the evaluation, research disciplines were distributed across eight broad clusters (e.g. "Mathematical, Information and Computing Sciences", "Physical, Chemical and Earth Sciences"). Quantitative indicators were identified and collected for the submitted outputs (including books and book chapters, conference publications, journal publications published in a journal included in the ERA Submission Journal List, recorded works, etc.), with focus on those indicators that "relate most closely to the quality of research outputs —- such as citation metrics and peer review" [Australian Research Council, 2015a].

In addition, peer review was used to inform the expert evaluation in certain disciplines (particularly humanities and social science disciplines), where it was felt quantitative indicators don't provide sufficient evidence of research quality. Peer review generally wasn't used in STEM disciplines. Because the focus of this section is on studying how different evaluation frameworks perceive and define research quality, we do not analyse the quantitative indicators used in ERA, but instead focus on the criteria used in the peer review evaluation. These criteria are *approach* and *contribution*, which are described as [Australian Research Council, 2015a]:

- **Approach** "is described as the approach taken in the group of outputs reviewed, potentially including reference to the methodologies, appropriateness of outlets/venues and discipline-specific publishing practices."

- **Contribution** "is described as the contribution of the group of outputs reviewed to the field and/or practice."

The provided definitions of approach and contribution are very broad. Furthermore, it can be seen from the description of "approach" that it contains two separate criteria: (1) approach taken in terms of preparing the outputs (methodologies, etc.), (2) approach taken in terms of publishing. We will further use these two criteria separately, as in our view publishing practices don't necessarily relate to research quality. However, an appropriate venue can potentially help to improve dissemination.

3.1.3 New Zealand

The New Zealand's national evaluation exercise, Performance-Based Research Fund (PBRF), is managed by the Tertiary Education Commission

(TEC) and was so far conducted in 2003, 2006 and 2012, with the next round scheduled for 2018 [Tertiary Education Commission, 2017].

Because the aim of the evaluation is, aside from increasing the overall quality of research, to support tertiary and postgraduate education [Tertiary Education Commission, 2013a], a significant portion of the evaluation is based on research degree completion and external research income (25% and 15%, respectively). The remaining 60% of the score is based on the assessment of performance of research staff, which is composed of evaluation of research outputs (weight of 70%), peer esteem (recognition of the staff member by peers such as through awards, fellowships, and panel participation; this part has a 15% weighting), and contribution to research environment (for example through supervision of students; this part constitutes 15%). Here we focus on the evaluation of research outputs.

The 2012 evaluation was based on peer review, which was conducted by 12 panels within 12 subject areas (including for example "Mathematical and Information Sciences and Technology", "Physical Sciences" and "Education"). The accepted research outputs included conference and journal articles, books, dissertations, software and design, and they were evaluated against a seven-point scale with descriptions provided for four of the seven points (so-called tie-points) [Tertiary Education Commission, 2013b]:

- **6 points:** Research characterised by "outputs that represent intellectual or creative advances, or contributions to the formation of new paradigms, or generation of novel conceptual or theoretical analysis and/or theories or important new findings with wider implications. In doing so it could indicate research that is exemplary in its field and/or at the leading edge and/or highly innovative. It would be expected to demonstrate intellectual rigour, imaginative

insight or methodological skill or to form a primary point of reference to be disseminated widely. A significant proportion of research outputs should be presented through the most appropriate and best channels. The research outputs would be likely to result in substantial impact or uptake. Such impacts could also include: product development, uptake and dissemination; or significant changes in professional, policy, organisational, artistic, or research practices."

- **4 points:** A publication representing a "significant research output that has generated substantial new ideas, interpretations or critical findings and that makes a valuable contribution to existing paradigms and practices. The research outputs generate new information or ideas and are well researched and technically sound. The EP typically includes research outputs that are presented in reputable channels considered as being at least at a middle level of excellence. The research is likely to contribute to further research activities and to have demonstrable impacts reflected in developments that may include: product development, uptake and dissemination; or changes in professional, organisational, policy, artistic, or research practices."

- **2 points:** Research characterised by "research activity (or developing research activity) and output that is based on a sound/justifiable methodology, and that makes a contribution to research within the discipline and/or to applied knowledge. This could be demonstrated by the production of research outputs that have been subject to quality-assurance processes."

- **1 point:** "Minimal evidence of research activity. The research outputs are assessed as having limited or no significance/impact, as contributing little or no additional understanding or insight in

the discipline/field, and/or as lacking in the appropriate application of theory and/or methods."

It can be seen the descriptions generally mention four broad themes: (1) research contribution and innovativeness, (2) methodological/technical soundness, (3) appropriateness and quality of publication channels, (4) impacts. However, in our view, while the first two criteria (contribution and methodological/technical soundness) can be seen as **aspects of research quality**, the latter two criteria (venues and impacts) are **evidence of research impact or importance** rather than aspects of quality. This is because the first two criteria directly influence/are manifested in the content of the evaluated research outputs, while the latter two don't have direct influence on the content, but can potentially exhibit different attributes as a consequence of a certain qualities of the research. Furthermore, the relation between venue and publication impact has been shown to be limited [Seglen, 1994, 1997].

3.1.4 Italy

In Italy, the agency responsible for assessing research quality is the National Agency for the Evaluation of Universities and Research Institutes (ANVUR). The Italian evaluation exercise, called VQR (Research Quality Evaluation), has so far been performed three times: (1) the first covered the period of 2001-2003, the second 2004-2010, and the last 2011-2014 [Ancaiani et al., 2015, Franceschini and Maisano, 2017]. Here we review the latest version of the exercise.

According to the call for participation [National Agency for the Evaluation of Universities and Research Institutes, 2015], 75% of the final score was based on a score for the quality of research outputs and the remaining 25% was based on other indicators (research funding, Ph.D.

programs, etc.). The evaluation of research outputs (namely journal contributions, scientific monographs, book contributions, such as chapters and conference proceedings, patents, and other outputs including data and software) was conducted using one or both of two methodologies: (1) bibliometric indicators (focusing mainly on citation counts and journal impact), (2) peer review. The evaluation was based on the following criteria [National Agency for the Evaluation of Universities and Research Institutes, 2015]:

- **Originality**, which is defined in the call for participation as "the level at which the research output introduces a new way of thinking in relation to the scientific object of the research, and is thus distinguished from previous approaches to the same topic".
- **Methodological rigour**, which is defined as "the level of clarity with which the research output presents the research goals and the state of the art in literature, adopts an appropriate methodology in respect to the object of research, and shows that the goal has been achieved".
- **Attested or potential impact** upon the international scientific community of reference, defined as "the level at which the research output has exerted, or is likely to exert in the future, a theoretical and/or applied influence on such a community also on the basis of its respect of international standards of research quality".

Similarly as in case of New Zealand's PBRF, the criteria used in the Italian national evaluation exercise could be divided into two groups: (1) criteria which directly reflect the quality of the published research (originality, methodological rigour), (2) criteria which provide an indication of importance (attested or potential impact).

3.1.5 Netherlands

The Standard Evaluation Protocol (SEP), which is the research evaluation system used to evaluation research in the Netherlands, is managed by the Royal Netherlands Academy of Arts and Sciences (abbreviated KNAW), the Netherlands Association for Scientific Research (NWO), and the Association of Universities in the Netherlands (VSNU) [Royal Netherlands Academy of Arts and Sciences, 2017]. SEP is performed in six-year cycles, the last of which was performed in years 2009–2015, with the next one scheduled for 2015–2021 [Royal Netherlands Academy of Arts and Sciences, 2009].

SEP consists of self-evaluation, external review, and site visits. The evaluation is conducted for whole institutes as well as for separate research groups or programmes, and is performed according to four main criteria: (1) quality of research (quality and scientific relevance of the research, leadership, reputation, organizational aspects, and PhD training), (2) productivity (in terms of inputs – staff and funds – and outputs – publications, dissertations, patents, etc.), (3) societal relevance (interaction and relevance to stakeholders or procedures, such as laws and regulations; also valorisation, i.e. making results available through products and services), (4) vitality and feasibility (ability to respond to change, management of projects, etc.). Here, we focus on quality of research, as in SEP, quality of research outputs falls within this criterion. Quality of research is one of five attributes of overall quality, and is described as follows [Royal Netherlands Academy of Arts and Sciences, 2009]:

> "**Quality and scientific relevance of the research:** Originality of the ideas and the research approach, including technological aspects; Significance of the contribution to the field; Coherence of the programme; Quality of the scientific

publications; Quality of other output; Scientific and technological relevance."

The document describing the 2009–2015 evaluation doesn't provide any description of the individual criteria (such as "Quality of the scientific publications"). One of the listed criteria, "Coherence of the programme", is related to the institute/research group rather than to the outputs. Otherwise, it can be seen the evaluation is broadly focused on (1) originality, (2) significance, (3) relevance of research, (4) quality of outputs (presumably methodological and technological rigour).

3.1.6 Summary

The quality criteria used in the five reviewed systems are summarised in Table 3.1. It can be seen there are three criteria which repeat across multiple systems: (1) originality/contribution (although these two concepts are not exactly the same, they are strongly related, and in the descriptions shown above originality is often explained in terms of contribution), (2) significance/impact/relevance to the field and to practice, (3) scientific and methodological rigour. One further criterion appears in one of the systems: appropriateness and quality of publication venues. As the latter criterion is specific to one system, here we focus on the former three. These are broadly described by the systems as:

- **Originality/contribution:** a creative/intellectual advance that makes a contribution to the field and state-of-the-art (such as new paradigms, theories, ideas, interpretations, methods, findings, problems, forms of expression), distinctive, or transformative work.
- **Significance/impact/relevance:** advancement of knowledge, thinking, skills, understanding, scholarship, and education; or applicab-

Table 3.1: Quality criteria used in different research evaluation systems.

Country	**System**	**Criteria**
United Kingdom	Resarch Excellence Framework (REF)	(1) Originality, (2) Significance, (3) Rigour
Australia	Excellence in Research for Australia (ERA)	(1) Methodological approach, (2) Appropriateness of venues (2) Contribution
New Zealand	Performance-Based Research Fund (PBRF)	(1) Contribution and innovativeness, (2) Methodological/technical soundness, (3) Appropriateness and quality of publication venues, (4) Impacts
Italy	Research Quality Evaluation (VQR)	(1) Originality, (2) Methodological rigour, (3) Attested or potential impact
Netherlands	Standard Evaluation Protocol (SEP)	(1) Originality, (2) Significance, (3) Relevance, (4) Rigour

ility to products, services, policies or other practical applications; scientific and technological relevance.

- **Rigour:** thoroughness in conducting the research, including appropriateness of methodology, clear description of statement of purpose, and coherence and consistency of analysis and arguments.

For simplicity, we will further refer to these criteria as originality, significance, and rigour. Although these definitions are very broad, they give us a better understanding of what aspects or dimensions of quality exist and how are they typically categorised. In the next few section, we will attempt to provide more detailed descriptions of these dimensions.

3.2 Journal peer review

Peer review is seen by the scientific community as a mechanism for controlling the quality and quantity of published research [Armstrong, 1997, Nature Neuroscience Editors, 1999, Kelly et al., 2014], and, regardless of its limitations [Teixeira da Silva and Dobránszki, 2015, Francois, 2015], it is generally considered to be the best available measure for filtering out good research from bad [Garfield, 2003, Bornmann and Daniel, 2005, Kreiman and Maunsell, 2011]. Here, we look at the criteria typically used in journal peer review for selecting manuscripts for publication. The focus of this chapter is not on studying peer review in itself, but solely on analysing the criteria used in journal peer review. The summary of this review is provided in Section 3.2.1.

Motivated by a highly visible case of a Korean researcher in the field of stem cell research, Dr. Hwang Woo Suk, who fabricated a series of experiments which appeared in high-profile journals, Bornmann et al. [2008] conducted one of the most extensive reviews of criteria used in journal peer review. They reviewed 46 studies that examined criteria used by editors and reviewers when selecting manuscripts for publication. Their aim was on understanding whether reviewers look for scientific misconduct when reviewing papers, and their study identified 572 different decision criteria. As the study represents a very extensive review, here we analyse their findings and compare them with a more recent study by Nedić and Dekanski [2016].

In their study, Bornmann et al. [2008] analysed 46 papers which examined editors' and reviewers' criteria for selecting manuscripts for publication. The authors collected 542 unique criteria from the 46 studies. In the next step, they developed a classification system with nine categories and assigned all 542 to one of the following categories: (1) relev-

ance of contribution, (2) writing/presentation, (3) design/conception, (4) method/statistics, (5) discussion of results, (6) relevance to the literature and documentation, (7) theory, (8) author's reputation/institutional affiliation, (9) ethics. Each criterion was also labelled "positive", "negative" or "neutral" depending on sentiment of the statement. For example, of three statements assigned to category "relevance of contribution", the criterion "the topic selected was appropriate" was assigned a "positive" label, "contains nothing new" was assigned a "negative" label, and the criterion "appropriateness of topic" was assigned a "neutral" label. Figure 3.1 shows a distribution of these three label types across the nine categories.

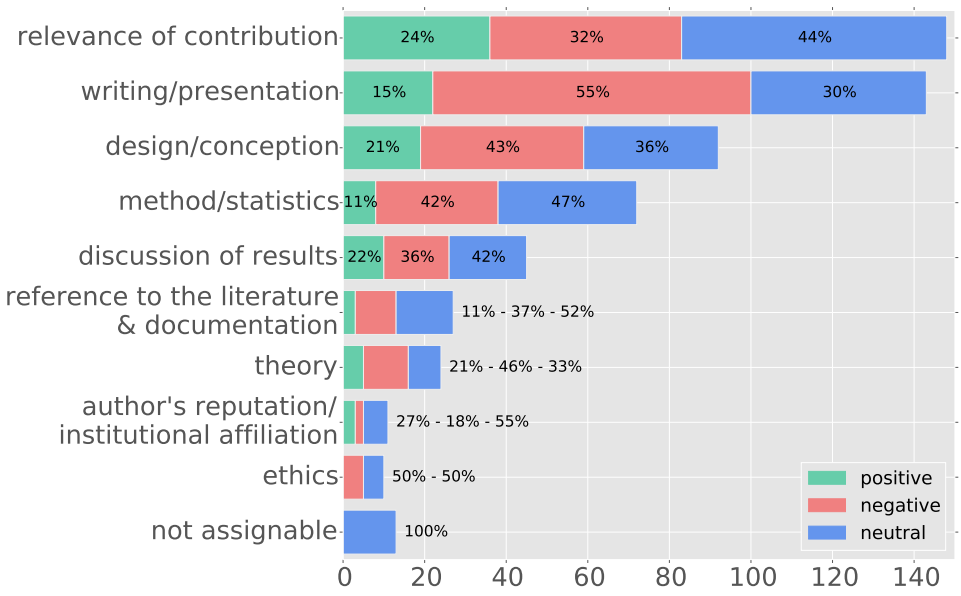

Figure 3.1: Distribution of decision criteria used by editors and reviewers for acceptance and rejection of journal manuscripts. Data from [Bornmann et al., 2008].

Each of the nine categories was further broken down into up to six subcategories and each criterion belonging to that category was assigned to one subcategory. For example, the subcategories for category "rel-

evance of contribution" were: (a) relevance of topic in general (such as whether the selected topic was appropriate, timely, important, relevant), (b) relevance of topic to scientific advancement (advancement of knowledge, topic pertinent to current research), (c) originality, newness (originality/novelty, creativity of ideas), (d) relevance of topic to journal (relevance to journal's focus, interest to readers), (e) contribution to practical progress (contribution to practice, usefulness of implications), (f) relevance of results (conclusive, complete results). For a complete list of subcategories and number of criteria belonging to each subcategory see [Bornmann et al., 2008].

Each but one of the categories identified by [Bornmann et al., 2008] can be assigned to one of the three main criteria presented in section 3.1: originality, significance, and rigour. Specifically, categories "ethics", "writing/presentation", "design/conception", "reference to the literature & documentation", "method/statistics", "discussion of results", and "theory" relate to rigour, and the category "relevance of contribution" relates partly to originality and partly to significance (depending on subcategory). The category "author's reputation/institutional affiliation" represents external evidence of potential quality and doesn't match any of the three criteria.

Interestingly, nearly all of the studies analysed in [Bornmann et al., 2008] mentioned several (on average three) criteria belonging category (1) relevance of contribution and category (2) writing/presentation. It can be seen in Figure 3.1 that a comparatively high proportion of criteria belonging to category (2) were negative. This would suggest issues related to writing or presentation of a publication are often criticised in the peer review process. This is also the case for category (3) design/conception. Some of the negative criteria mentioned in [Bornmann et al., 2008] belonging to category (2) include incoherent, bad tone, insufficiently de-

scribed subjects, tables/figures need clarification, lack of organization. Negative criteria belonging to category (3) included conceptual basis for study poor or incomplete, inadequate research design, sample too small or biased.

A recent publication by Nedić and Dekanski [2016] presented results of a survey on the importance of peer review criteria according to reviewers of the Journal of the Serbian Chemical Society (JSCS). The criteria studied by Nedić and Dekanski [2016] are (1) scientific contribution and originality, (2) clarity and conciseness, (3) length, (4) conclusions completely supported by results, (5) references, (6) quality of illustrations, (7) nomenclature in accordance with SI and IUPAC, (8) language (grammar and syntax). These specific criteria were formulated by the editors and sub-editors of JSCS. Figure 3.2 shows the distribution of answers chosen by the respondents.

The criteria studied by Nedić and Dekanski [2016] represent a subset of the assessment criteria identified by Bornmann et al. [2008]. Specifically, two of Nedić and Dekanski's criteria correspond to two of Bornmann et al.'s categories: criterion "scientific contribution and originality" matches category (1) "relevance of contribution" and criterion "references" matches category (6) "relevance to the literature and documentation". Five of Nedić and Dekanski's criteria ("language", "nomenclature in accordance with SI and IUPAC", "quality of illustrations", "length", "clarity and conciseness") match one of the subcategories of category (2) "writing/presentation". Finally, the criterion "conclusions completely supported by the results" matches one of the subcategories of Bornmann et al. category (5) "discussion of results". The following categories from [Bornmann et al., 2008] are not covered by Nedić and Dekanski [2016]: (3) "design/conception", (4) "method/statistics, (7) "theory", (8) "author's reputation/institutional affiliation", (9) "ethics". This matches

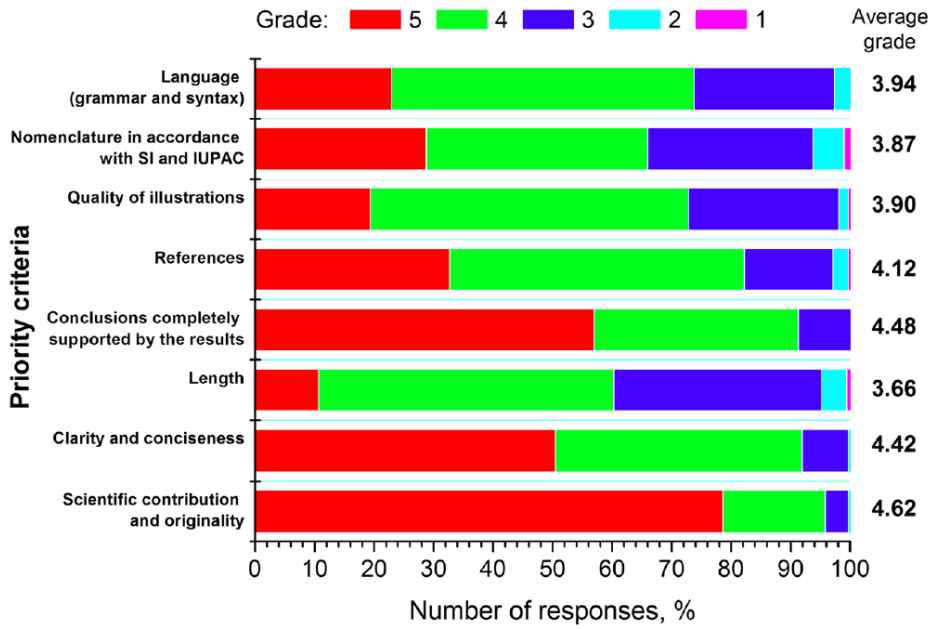

Figure 3.2: Grading of acceptance criteria according to reviewers of JSCS. The grading scale ranges from not important (1) to extremely important (5). The selected grades are expressed in % of the total number of responses. Source: [Nedić and Dekanski, 2016]. Reprinted by permission from Springer Nature: Springer Netherlands Scientometrics 107: 15, Priority criteria in peer review of scientific articles, Olgica Nedić and Aleksandar Dekanski, Copyright Akadémiai Kiadó, Budapest, Hungary 2016, advance online publication, 1 January 2016 (doi: doi.org/10.1007/s11192-016-1869-6.)

Bornmann et al.'s findings, as not all of the studies analysed in their paper mentioned these criteria.

Of the criteria studied by Nedić and Dekanski [2016], three were graded as very important by more than half of the reviewers: "scientific contribution and originality," "conclusions completely supported by the results," and "clarity and conciseness". The least important criterion was length, and the criteria belonging to Bornmann et al.'s category (2) "writing/presentation" were generally seen as less important. This

would suggest that although criteria related to qualities of reporting are frequently considered, they carry less weight than criteria related to qualities of the research being reported.

3.2.1 Summary

Many publications have studied reviewers' and editors' criteria for selection of articles for publication. Bornmann et al. [2008] have conducted an extensive review of 46 such studies, which we have analysed in this section and compared to a recent study by Nedić and Dekanski [2016]. Nedić and Dekanski [2016] have collected importance ratings of selected criteria for manuscript publication from editors and reviewers of their journal. The comparison has shown that research contribution is generally considered the most important aspect of publication quality, and that criteria related to writing and presentation get frequently mentioned in journal peer reviews (especially with negative sentiment) but are not perceived as very important by most reviewers. Rigour-related criteria (conclusions supported by results) also ranked fairly high in terms of importance.

3.3 Studies of research quality and influence

Two previous studies have looked closer at the actual concept of research quality (or, in the second case, impact): [Andersen, 2013] and [Sternberg and Gordeeva, 1996]. In this section we review these studies. A summary of our findings is presented in Section 3.3.1.

Sternberg and Gordeeva [1996] have studied the opinions of psychologists on what makes an article influential in psychology. Although their

focus is on influence rather than quality, the two concepts are related and influence is often seen as a dimension of quality. The aspects of influential research identified by Sternberg and Gordeeva [1996] will therefore likely also have an effect on the perceived research quality. Sternberg and Gordeeva [1996] have approached the study in two steps. They first collected a set of statements related to research impact in the field of psychology. The statements were collected from 20 psychologists and the final list, after removing duplicates, contained 45 statements.

In the second step they created a questionnaire asking psychologists to provide ratings of the importance of the statements collected in the first step. In total, 252 individuals returned a completed questionnaire. The statements were ranked on a scale from 1 (not at all important) to 6 (extremely important). The highest ranking criteria were: "makes an obvious contribution to psychological knowledge, adding something new and substantial", "presented results are of major theoretical significance", and "presents a useful new theory or theoretical framework". Least important among the listed criteria were "includes concrete examples", "provides evidence that supports an existing influential psychological theory", and "contains useful implications for professional practice". The ranking of importance of different statements reported in [Sternberg and Gordeeva, 1996] is largely in agreement with the frameworks and studies discussed in the previous two sections. For example, a common observation made across all previously mentioned publications is the perceived importance of research contribution for publication quality.

The authors have also conducted a principal component analysis of the complete correlation matrix and identified six factors which they interpreted as (1) quality of presentation (although this factor accounted for most variation in the data, the related criteria were ranked third in importance), (2) theoretical significance (the related criteria ranked highest

according to importance), (3) practical significance (these statements had the lowest mean importance), (4) substantive interest (whether the topic is interesting and captures reader's interest, these criteria were ranked fourth in importance), (5) methodological interest (new or interesting methodology or experimental paradigm or surprising results, these criteria ranked fifth), (6) value for future research (implications and/or recommendations for future research or for understanding of the field, these criteria ranked as second most important). This is again in line with previously discussed findings, as the two highest ranking factors ("theoretical importance" and "value for future research") contained most criteria related to research contribution and originality (e.g. new theory of theoretical framework, better explanation of existing phenomena, debunks an existing theory, implications/recommendations for future research). Interestingly, "practical significance" (criteria related to applicability of the research in practice) was the lowest ranking factor in terms of importance.

A similar study was conducted by Andersen [2013], whose focus was on identifying dimensions of research quality in medicine. Similarly as in Sternberg and Gordeeva [1996], the author has first conducted an interview study to collect a set of statements about research quality criteria relevant in the medical field, which yielded a list of 32 criteria. An online survey was then constructed to quantify the collected criteria. In total 279 individuals responded the survey, which included researchers in academia and industry as well as healthcare practitioners. The respondents were asked to rank the questions on a scale from 0 (completely disagree) to 5 or 10 (completely agree).

Once the responses were collected, factor analysis was performed to group and help to narrate the criteria. The analysis has identified six factors, each composed of one or several criteria [Andersen, 2013]: (1)

journal prestige (quality, prestige and effect from journal impact factor), (2) clinical guidelines (use and meaning of clinical practice guidelines), (3) use of references, (4) method section (the length of the method section), (5) subjective quality (peer review and clinical relevance), (6) basic to applied (purpose of basic research and clinical relevance), (7) prominence of the author, (8) citation meaning, (9) citation quality, (10) innovation stunt (whether peer review stunts innovative and groundbreaking research), (11) scepticism (whether there is an overflow of journals and scepticism towards clinical practice guidelines), (12) propriety (publishing of negative findings and use of certain golden standard methods).

It can be seen that while some of the factors are valid across disciplines, some are specific to the medical research field (such as "clinical guidelines"). Furthermore, several of the factors which are related to external factors rather than to the manuscript itself (journal prestige, peer review, prominence of the author, citations, propriety) could be seen as indicators of possibly high quality research rather than as criteria directly influencing publication quality. In fact, out of all factors studied by Andersen [2013], the factors directly influencing the manuscript are clinical guidelines, use of references, method section, basic to applied, and propriety. The respondents agreed about the need to publish negative findings (related to factor "propriety") and the purpose of basic research (which should aimed towards improving overall health). The respondents were divided into two groups with regards to the length of the method section, with one group claiming the length shouldn't be limited to allow for reproducibility and the other group claiming the length should be limited to improve readability.

3.3.1 Summary

Previous research has studied the opinion of practitioners working in different scientific fields on the concept of research quality and influence. Here we have analysed two such studies [Sternberg and Gordeeva, 1996, Andersen, 2013]. Sternberg and Gordeeva [1996] have studied the perspective of psychologists on what makes an article influential in the field of psychology. Although their study was focused on influence rather than quality, many of the factors they identified were in line with the findings reported by other studies discussed earlier in this chapter and were strongly related to quality. Their findings suggested research contribution and theoretical significance are among the most important factors. On the other hand, Andersen [2013] has focused specifically on research quality; however, his findings were in many cases not directly related to publication quality, but rather to external evidence such as citations and impact factors.

3.4 Survey of researchers' perspective

In the previous sections we have described several national research evaluation exercises and studies concerned with research publication quality, influence, and peer review criteria. This has provided us with statements related to specific characteristics of research publications related to research quality. These statements typically relate to one of three main criteria: (1) the publication's contribution, innovativeness and originality, (2) significance and relevance of the research to the field and to practice, (3) scientific and methodological rigour.

We have conducted an online survey to (a) gain a better understanding of the importance of the specific characteristics of research publications related to quality (are there any characteristics which are gener-

ally considered very important? These would be a priority for further studies and development of new research evaluation metrics), and (b) analyse the relationship of the three main criteria and their relation to quality (can a publication still be considered of high quality if it lacks rigour/significance/originality?). This section describes the survey and provides an analysis of the responses. First, in Section 3.4.1 we describe the format of the survey and present summary statistics describing our respondents. In Section 3.4.2 we present and analyse the results of the survey. We summarise our findings in Section 3.4.3.

3.4.1 Data collection

Our survey was inspired by the studies by Sternberg and Gordeeva [1996] and Andersen [2013]. Both studies were done in two phases, the goal of the first phase was on generating statements about aspects of quality and the second on verifying and ranking the collected aspect. We have constructed our survey in a similar way. It was composed of four parts:

1. questions about the respondents background and experience (their discipline and seniority),

2. a set of open-ended questions asking the participants to list publications they think are of high quality and to specify why they think so,

3. characteristics of research publications related to originality, significance, and rigour which the participants were asked to rank on a scale from 0 (statement not indicative of a given criterion) to 10 (extremely indicative),

4. questions about the relation between originality, significance, rigour, and quality.

We have collected statements on aspects of research quality identified by the studies mentioned in the previous sections and assigned these aspects to the three main broad criteria identified in the previous sections: originality, significance, and rigour. Questions about the importance of these statements formed the third part of the survey. The aim of the second part of the survey (the open-ended questions) was on understanding whether there are any important characteristics which were omitted in the third part. The complete survey together with the invitation email can be seen in Appendix A.

The survey was sent to academic staff and research students from all faculties of the Open University (to 1,409 people in total). The reason why we contacted Open University researchers is because research at the Open University covers many disciplines, and because it is the largest university in the UK. We were therefore able to get a significant sample spanning multiple disciplines. Within two months we received 105 responses, which represents a 7% response rate.

In order to define the respondents' professional background, seniority, and publication record, they were first asked three questions: (1) which research area they feel most associated with, (2) how many years ago did they received their PhD, (3) how many publications they authored during their career. The list of disciplines presented to the respondents matched the units of assessment used in the latest Research Excellence Framework (REF) [Research Excellence Framework, 2014a]. We have selected this classification because UK researchers are familiar with it. Figure 3.3 shows the number of responses received per each of the main REF panels [Research Excellence Framework, 2014a].

Out of the 105 respondents, 11 have selected "Other" instead of one of the predefined areas. The explanations provided for the selection mostly mentioned multidisciplinary research (e.g. "Computer Science AND Edu-

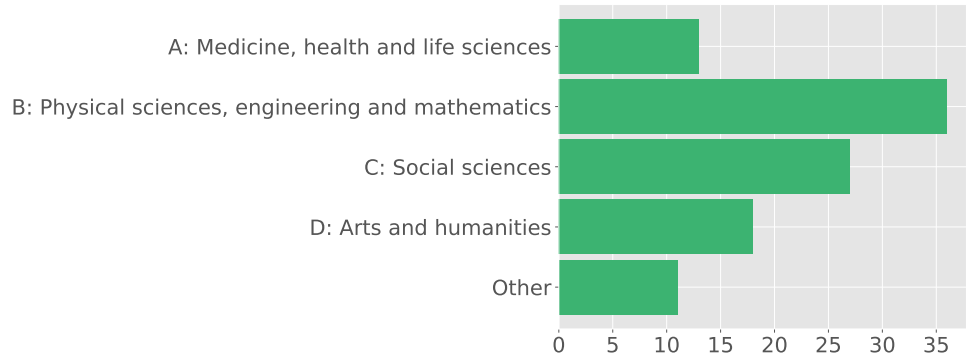

Figure 3.3: Number of responses received per each of the main REF panels.

cation", "Learning Analytics", "Mathematics Education"). The respondents were also asked to provide specific areas of interest; however, as these areas are more detailed and there is little overlap between them we haven't used these in our analysis.

Next, the respondents were asked to provide a number specifying how long have they held their PhD (or "0" in case they didn't have a PhD at the time of filling the questionnaire). Figure 3.4 (right) shows the number of respondents according to the number of years since they received their PhD. For the number of authored publications the respondents were given 6 options ("5 or less", "6-15", "16-25", "26-50", "51-100", "More than 100"). Figure 3.4 (left) shows the number of respondents in according to their publication record. Table 3.2 shows a comparison of the two statistics.

3.4.2 Survey results

Open-ended questions

As explained in the previous section, the second part of the survey consisted of open-ended questions asking the participants to think of public-

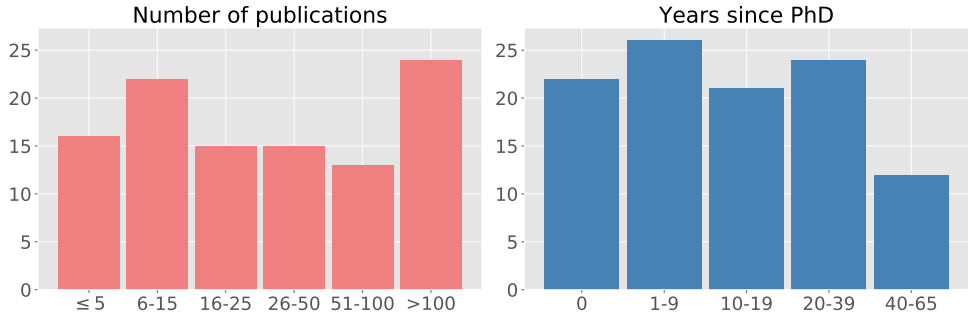

Figure 3.4: Number of respondents in terms publication record (left) and seniority (right).

Table 3.2: Comparison of seniority and publication record of the respondents.

		Years since PhD					
		0	**1-9**	**10-19**	**20-39**	**40-65**	**All**
# of publications	**$\leq$ 5**	10	4	2	0	0	16
	6-16	5	9	6	2	0	22
	16-25	3	5	4	3	0	15
	26-50	1	6	3	3	2	15
	51-100	1	2	3	4	3	13
	>100	2	0	3	12	7	24
	All	22	26	21	24	12	105

ations they consider of very high quality and to explain why they think these publications are of high quality. This part of the survey was inspired by studies by Sternberg and Gordeeva [1996] and Andersen [2013]. The goal was to introduce the respondents to the problem studied in the survey and to understand whether the latter parts of the survey correctly addressed the most important aspects of quality. Moreover, the list of high quality publications provided by the respondents can serve as a dataset for further studies comparing these publications with a background population.

To analyse the answers, we have done the following. As the respondents were allowed to enter text of any length and any number of reasons, we have first split each answer into separate statements. For example, one respondent provided the following answer to the question asking why they consider the listed publications to be of high quality: "Good and deep explanation of methodology; clear results; easy to reproduce." We have split this answer into three statements: (1) "good and deep explanation of methodology", (2) "clear results", (3) "easy to reproduce". Next, we have merged similar statements. For example, statements "they make a substantial contribution to the fields of economics and finance" and "they bring an interesting contribution to the body of knowledge" were merged into "contribution to the field". Finally, we have grouped the statements according to a general high-level category they were related to.

The analysis of answers to the question "Why do you consider the publications you listed in the previous step to be of high quality?" has revealed 328 statements, which were collected from 86 completed answers (19 respondents out of the 105 in total did not provide an answer to this question). After merging similar statements, we were left with 252 unique statements. We have assigned each of these statements to one of the following categories: (1) originality, (2) rigour, (3) significance, (4) external evidence (statements mentioning the author, publication venue, or opinion of peers), (5) other (statements which couldn't clearly be assigned to any of the other categories). Out of the total 252 unique statements, 63 were assigned to the category "originality", 73 to the category "rigour", 43 to the category "significance", 27 to the category "external evidence", and 44 to the category "other". Figure 3.5 shows frequency of statements and number of new unique statements added per answer/respondent. The most frequently mentioned aspects were

"number of citations" (mentioned 9 times), "innovative", "well written" (both mentioned 7 times), "clarity of presentation" (mentioned 6 times), "contribution to the field", "ground breaking", "new ideas", "rigorous", "peer review" (all mentioned 4 times).

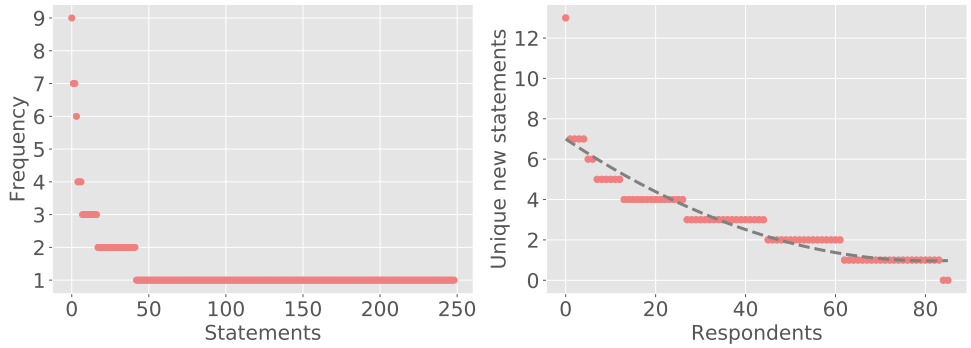

Figure 3.5: Frequency of statements (left), and number of new unique statements added by participant (right). In both plots, x-axis is sorted by frequency/count.

Most of the statements belonging to the category "originality" were related to the type of contribution the publication made (for example "solved an outstanding problem", "clarifies aspects of the field", "opened path for research in the area", "fills gap in literature"), while the rest of the statements were related to originality/novelty of the publication (for example "first to investigate a new topic", "first to answer a question", "novel finding"). We have further split the category "rigour" into two subcategories: (1) statements related to quality of methodology, use of theory, evaluation, analysis, and experimentation (how the research is done, these statements included for example "thorough evaluation" and "transparent methodology"), (2) and statements related to the quality of writing and presentation (how the research is reported, these statements included for example "choice of methodology explained" and "long introduction"). These subcategories were assigned 44 and 30 statements,

respectively. We have created a separate category for statements about external evidence of publication significance, as these statements don't directly describe the type of significance, impact, or relevance of the publication. The category "significance" was therefore assigned statements, such as "used in teaching", "clinical outcome" and "applicable in practice", while the category "external evidence" contained statements related to the publication venue ("venue acceptance rate"), type and number of citations ("cited by prominent authors"), and other factors ("Nobel prize"). Finally, the category "other" contained statements we were unable to assign to any of the other categories, particularly statements related to interestingness of the topic of the publication (e.g. "addresses a well-established field"). The complete set of statements assigned to each of the categories can be found in Appendix A.

Ranking aspects of originality, significance and rigour

The third part of the survey was devoted to analysing the importance of specific characteristics of research publications related to originality, significance, and rigour. This part of the survey was formulated as a set of statements and the respondents were asked to specify how indicative is each of the statements of originality, significance, or contribution on a scale from 0 (not at all) to 10 (extremely). The list of statements was produced by combining relevant statements identified by Sternberg and Gordeeva [1996] and Andersen [2013], and complementing the list with our own statements where we felt an important characteristic was missing. The list contained 16 statements related to originality, 36 statements related to rigour, and 22 statements related to significance.

Table 3.3 lists the complete set of aspects of originality as well as the mean rating and standard deviation of the rating, ranked by mean rating in descending order. The mean values show that the aspects that are the

most indicative of publication originality are related to providing new knowledge, ideas, theories, data, etc., while supporting existing theories, combining and applying known methods, and providing generalisations ranked lowest. Interestingly, providing evidence that fails to support an existing theory ranked higher than providing evidence in support of an existing theory, possibly because the former may indicate the need for change or for further development of the theory.

Table 3.3: Basic statistics on aspect ratings for the aspects related to originality.

	Aspect	**Mean**	**SD**
1	Provides new knowledge	7.48	1.82
2	Provides new ideas	6.88	2.02
3	Presents a new theory or theoretical framework	6.80	2.23
4	Presents a new viewpoint on a problem	6.75	2.04
5	Opens up a new problem (research question) for investigation	6.67	2.00
6	Presents a new method (methodology, experiment, test, technique, treatment, etc.)	6.65	2.27
7	Integrates many different areas of data previously thought to be unrelated	6.46	2.11
8	Connects and integrates work from multiple disciplines	6.37	2.32
9	Provides new data/resources enabling further research	6.33	2.38
10	Clarifies existing problem(s)	5.96	2.06
11	Provides evidence that fails to support an existing theory	5.93	2.58
12	Integrates into a new, simpler framework data that had previously required a complex and possibly unwieldy framework	5.89	2.19
13	Combining known methods in a new way	5.68	2.06
14	Applying known methods to a known problem for the first time	5.66	2.29
15	Provides evidence that supports an existing theory	5.62	2.46

	Aspect	Mean	SD
16	Contains generalisations, which are clearly stated, confirmed	4.94	2.40

The standard deviations show where there was the most and the least agreement among respondents. Both of the two highest ranking aspects were also among the aspects with the highest agreement. On the other hand, the aspect regarding evidence that fails to support an existing theory had the highest disagreement, and the two lowest ranking aspects overall came second and third in terms of disagreement among respondents.

The complete set of aspects related to rigour is shown in Table 3.4. Clearly stated and well-conceptualised problem, and well-explained and sound methodology were the highest ranking aspects as well as the aspects with the highest agreement among respondents. Somewhat surprisingly, the respondents disagreed the most about testing results for statistical significance and about reproducibility. Testing results for statistical significance was one of the lowest ranking aspects, while reproducibility ranked in the middle in terms of how indicative it is of rigour. The disagreement among respondents regarding these two aspects could be attributed to differences between disciplines. While for some disciplines (such as psychology) statistical testing is an important part of results analysis, for other disciplines (for example computer science) statistical testing is not utilised as much.

Table 3.4: Basic statistics on aspect ratings for the aspects related to rigour.

	Aspect	Mean	SD
1	The problem is clearly stated and well-conceptualised	7.43	1.99
2	If a new methodology is introduced, it is explained in enough detail	7.34	2.07

	Aspect	Mean	SD
3	If a new methodology is introduced, it is sound	7.19	2.00
4	The publication describes how the results were obtained	7.15	2.20
5	The publication objectively discusses the limitations of the results	7.13	2.29
6	The results are valid	6.92	2.62
7	Sources are cited for their importance and relevance (rather than collegiality, venue impact, etc.)	6.83	2.39
8	The results are discussed thoroughly (considering different interpretations and extreme cases)	6.81	2.35
9	The publication provides substantial and convincing evidence for proving or disproving the hypothesis	6.72	2.46
10	The publication presents the purpose and motivation for tackling the problem	6.67	2.42
11	The hypothesis is clearly stated	6.63	2.44
12	The publication discusses the contribution and importance of the results	6.62	2.28
13	The methodology selection matches the hypothesis and the data	6.59	2.47
14	The results interpretation is unbiased and unambiguous	6.55	2.86
15	The publication contains a description of the data collection	6.44	2.64
16	The experiment is described in enough detail to be reproducible	6.43	3.13
17	Clear and concise conclusion	6.42	2.34
18	Keeping the writing to the point	6.37	2.33
19	Clear, concise and grammatically correct language	6.24	2.57
20	Consistent writing	6.20	2.47
21	The publication presents valid but negative results	6.07	2.78
22	The literature review mentions in which way the paper makes a contribution to the field	5.98	2.51
23	The publication presents a proof of the results	5.89	2.95
24	The data involve a sufficient number of cases (data, samples, events, patients etc.)	5.83	2.83

	Aspect	Mean	SD
25	Clear and concise abstract	5.82	2.59
26	Contains implications for future research	5.81	2.53
27	Is easily understandable	5.80	2.94
28	Unbiased tone	5.72	2.66
29	The literature review section mentions all important relevant studies	5.54	2.58
30	The results are checked for statistical significance	5.40	3.20
31	The writing attracts and keeps attention	5.35	2.97
32	The paper is of an adequate length given the problem	5.32	2.90
33	The data used in the experiment are publicly shared and accessible	5.27	3.07
34	The publication builds on previous research	5.25	2.59
35	Contains recommendations for further research	5.18	2.71
36	The publication uses a well-established methodology	4.23	2.69

Table 3.5 shows a ranked list of aspects of publication significance. The aspects related to significance ranked on average lowest compared to aspects related to originality and rigour (average rank of 4.73 compared to an average rank of 6.26 for originality and 6.20 for rigour). Only two aspects of significance obtained a mean rank of 6 or higher: causing a significant knowledge shift and topic importance. Topic importance was also the aspect with highest agreement among respondents. Interestingly, the aspect with the second highest agreement was related to topic popularity. In this case, the respondents agreed topic popularity is not a very important aspect of significance. The lowest ranking aspects were receiving media coverage, resulting in a patent, and, somewhat surprisingly, resulting in a product or a service[1]. Conference and journal related

[1]We note that this aspect could possibly be skewed by the surveyed population (academics), and different groups (such as government employees or funders) might have answered differently.

aspects also ranked fairly low. Of the aspects related to citations, receiving citations within the publication's area ranked higher and had higher agreement between respondents than receiving many citations. It would therefore seem the respondents felt being recognised by the specialised area is more important than number of citations.

Table 3.5: Basic statistics on aspect ratings for the aspects related to significance.

	Aspect	**Mean**	**SD**
1	Results encouraged a significant knowledge shift	6.66	2.55
2	Topic is important	6.59	2.02
3	Further research builds on the results	5.93	2.47
4	Received citations within its specialised area	5.92	2.63
5	Is criticised or scrutinised by further research	5.65	2.46
6	Influenced professional practice (policies, recommendations)	5.60	2.97
7	Has been publicly acknowledged by the research community	5.51	2.78
8	Further research mentions the results	5.41	2.47
9	Received many citations	5.39	2.81
10	Received citations from outside of its area/field	4.87	2.82
11	Has been read by a significant number of people (e.g. as measured by downloads, views, bookmarks, etc.)	4.84	2.94
12	Has provided societal benefits (economic, social, etc.)	4.82	3.19
13	Has been published in a high-impact journal	4.76	2.81
14	Influences multiple disciplines	4.61	2.89
15	Is applicable in many areas	4.60	2.91
16	Has been presented at a high esteem conference	4.31	2.82
17	Has received funding as a result of the research	3.66	2.78
18	Topic is popular	3.29	2.11
19	Has generated public interest (e.g. as measured by tweets, non-academic invited talks, blog mentions, etc.)	3.24	2.79
20	Has resulted into a product or service	3.06	2.81

	Aspect	Mean	SD
21	Has resulted in a patent	2.67	2.66
22	Has resulted in media coverage (e.g. news coverage, etc.)	2.52	2.38

Relation between originality, rigour, significance and quality

The fourth and final part of the survey was focused on analysing the relation of originality, rigour, and significance to overall publication quality. This part of the survey consisted of a set of statements and the respondents were asked to specify how much do they agree with these statements on a scale from 1 (disagree) to 5 (agree). The complete set of statements studied in this part of the survey, along with mean rating and standard deviation, is presented in Table 3.6.

Table 3.6: Basic statistics on the relation of originality, rigour and significance to quality.

	Statement	Mean	SD
1	High quality research publications present rigorous research.	3.08	1.05
2	High quality research publications present original/novel research.	2.74	1.14
3	A low rigour research publication cannot be of high quality.	2.70	1.28
4	High rigour research publications are of high quality.	2.65	1.11
5	Significant research publications are of high quality.	2.57	1.12
6	High-quality research publications have higher significance.	2.44	1.21
7	Publications providing novel/original ideas are of a higher quality.	2.35	1.19
8	High significance of a research publication is an evidence of its quality.	2.18	1.26
9	The quality of a research publication is independent of its originality/novelty.	2.01	1.29
10	The level of significance of a research publication is independent of its quality.	1.74	1.29

	Statement	Mean	SD
11	A research publication lacking originality/novelty cannot be of a high quality.	1.73	1.30
12	The quality of a research publication is independent of its rigour.	1.01	1.03

Several things can be observed from the results. First, it seems the respondents perceived rigour and publication quality as strongly related. The respondents consistently agreed with the statement that high quality research publications present rigorous research, and that rigorous research publications are of high quality. At the same time, they consistently disagreed with the statement that the quality of a research publication is independent of its rigour, and agreed that a publication of low rigour cannot be of high quality. Figure 3.6 shows distribution of the responses for each of the statements, with the four statements related to rigour listed at the top of the figure.

The respondents also agreed with the statement that high quality research publications present original research. However, in this case, they didn't think a publication lacking originality cannot be of high quality. This would suggest that unlike in the case of rigour, originality is not perceived as a necessity for a publication to be of high quality; however, high quality publications are to a certain degree expected to be original. Finally, the respondents largely disagreed with the statement that the level of significance of a research publication and its quality are not related, and more than half of all respondents agreed with the statement that significant research publications are of high quality. It therefore seems that publications that became highly significant are presumed to be of high quality. The respondents didn't agree nor disagree with the remainder of the statements.

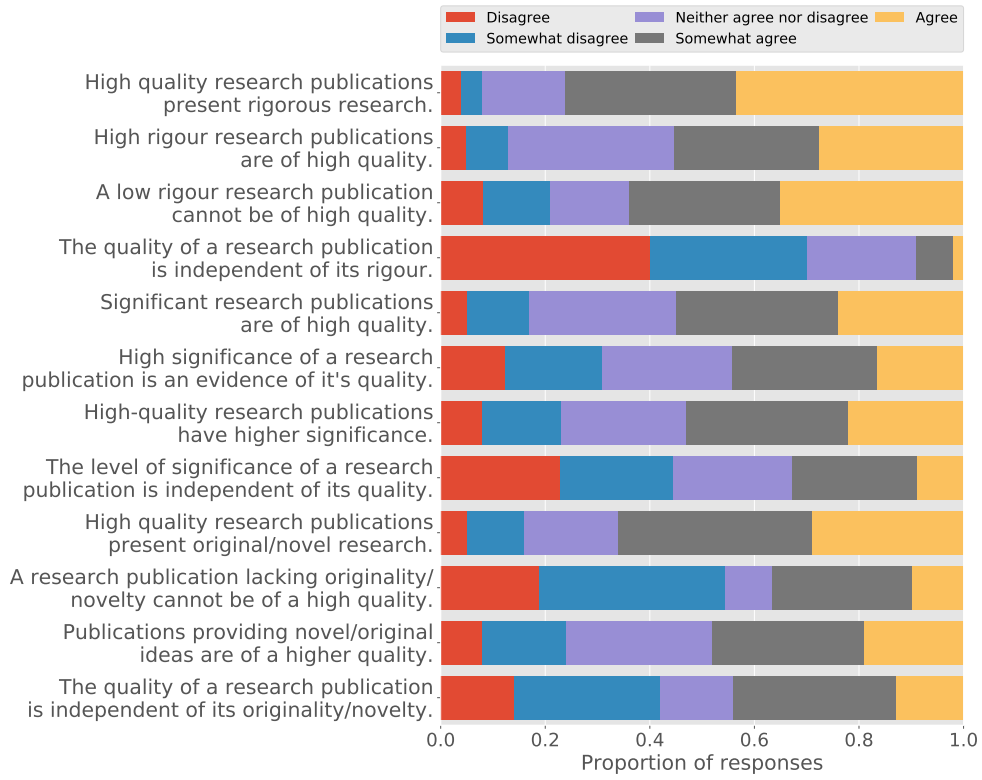

Figure 3.6: Grading of statements on the relation between publication quality and originality, rigour and significance.

3.4.3 Summary

In this section we have reported the results of an online survey in which we investigated researchers' view of research publication quality. As we have shown earlier in this chapter, research quality is typically described in terms of three main criteria: originality (the contribution the publication/research provided), rigour (how well was the research performed and the publication written), and significance (what/who did the research/publication affect). We have collected a set of aspects related to the three main criteria and asked the respondents to rank the aspect in terms of how important each aspect is for the relevant criterion. The survey revealed that when it came to originality, the respondents looked for a novel contribution (such as new knowledge, ideas, theory and data),

when it came to with rigour, the respondents thought clear problem statement and methodology selection and description were the most important aspects, and when it came to significance, causing a knowledge shift and enabling further research were among the most important aspects. We also investigated the relation between the three main criteria to overall publication quality. Overall, the respondents rated rigour as strongly related to publication quality. The analysis also revealed high quality publications were viewed as presenting original research.

3.5 Conclusions

This chapter provided an overview and analysis of criteria typically used to evaluate the quality of research and research publications, which addressed the following question: "What is research publication quality and what factors influence it?" We have approached the question in two steps. We have reviewed the criteria used in several national research evaluation exercises and in journal peer review. We have also reviewed two studies focused on identifying the dimensions of publication impact and quality. We have seen that across the different frameworks and studies, research publications are typically evaluated in terms of three broad criteria: (1) originality (the original contribution the publication provided), (2) rigour (how well was the research performed and the publication written), and (3) significance (what/who did the publication affect).

To understand which specific factors influence these three criteria, we conducted an online survey which was answered by 105 university researchers from different disciplines. Among other things, we found that statements related to originality and rigour ranked on average fairly high, and higher than statements related to significance. Our respondents viewed particularly rigour as strongly related to publication quality.

Originality was also viewed as related to quality, but to a lesser degree. The reason for this might be that rigour may be easier to judge than originality. This is because strong knowledge of the field may not be necessary to be able to judge a publication according to its rigour. Originality, on the other hand, may require prior knowledge. We have found a similar pattern with regard to rigour and originality in our literature review, where we observed that journal editors tend to often mention rigour in their reviews. However, in general, aspects related to research contribution were viewed as the most important criteria for publication acceptance by journal editors.

Chapter 4

Dataset and methods for research metrics evaluation

> *Everyone is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid.*
>
> – Albert Einstein

In the previous chapter we have studied the concept of publication quality from several different perspectives, which gave us a better understanding of which specific characteristics of research publications are typically seen as related to, or that are indicative of quality. This enables us to focus on specific publication characteristics when developing new methods. However, before developing new indicators and metrics for use in research evaluation, it is necessary to understand how can these new metrics be evaluated, i.e. how do we know these metrics work well and measure what was intended. This chapter addresses this question, that is:

> RQ2: *How can we evaluate the performance of metrics used in research evaluation for assessing the quality of research*

publications?

In order to be able to evaluate the performance of an indicator or a metric, two things are typically needed:

- A sample of research publications to test the metric on.
- A ground truth or reference data to compare the metric with in order to obtain a performance measurement. This could be human judgement (peer review), or results from another metric known to work well.

In this chapter, we review both publication datasets and evaluation approaches typically used for evaluating research metrics. Based on this review, we propose a new method, complementary to the existing evaluation approaches, and build a reference set which can be used for validating research metrics. We describe how this reference set was built, and, to ensure that it is suitable for this task, analyse several overview statistics describing it. Furthermore, we review and analyse the Microsoft Academic Graph, a new dataset of research publications which was recently released to enable research in mining scholarly publications, and which, due to its dense citation network and comprehensive metadata, is a promising dataset for scientometric research. A number of recent initiatives and reviews, including the Metric Tide Report (Chapter 2) mentioned the importance of openness and transparency of data and methods. The Microsoft Academic Graph provides such open resource.

This chapter is organised as follows. First, in Section 4.1, we review the existing openly available datasets of research publications which can be used to study research evaluation methods. In Section 4.2 we provide detailed analysis of a new dataset, the Microsoft Academic Graph, with focus on the applications of this dataset to research evaluation and related

areas. In Section 4.3 we discuss approaches which are typically used to analyse the performance of research evaluation metrics. Based on this review we develop a new, complementary dataset which can be used to evaluate the performance of research metrics. We discuss how this dataset was collected and present overview statistics of the dataset in Section 4.4. In Section 4.5 we summarise conclude our findings and effort presented in this chapter.

4.1 Research publication datasets

In this section we briefly review ten research publication datasets which can be used for scientometric research. A number of studies has previously compared the major citation indices, Clarivate Analytics Web of Science (WoS), Scopus, and Google Scholar, e.g. [Falagas et al., 2008, Fiala, 2011, Harzing and Alakangas, 2016]. While these citation indices are generally considered to be among the largest and most comprehensive, the downside is the difficulty of accessing their data – WoS and Scopus are commercial, and Google Scholar does not offer an API or bulk downloads. Reviews, comparisons and studies of other publication datasets are scarce. One such study has compared three additional datasets aside of the three main citation indices [Fiala, 2011], although with focus on Computer Science. However, knowing which datasets exist and being aware of their characteristics is important for understanding which datasets are suitable for which tasks. Therefore, in this section we provide a brief review and comparison of existing publication datasets. We focus on aspects important for research analysis and evaluation, such as multi-disciplinarity, and whether they contain citations and publication full texts. The aim of this review is not to be exhaustive in terms of inclusion of all known publication datasets, but to provide an overview

of some of the best known datasets, and their strengths and limitations. The datasets reviewed in this section were selected according to the following criteria:

- The dataset has to be publicly available to the research community. This requirement excludes both major databases, Clarivate Analytics Web of Science and Elsevier Scopus, from our study, as these are both commercial.
- It should offer a way to programmatically download data, such as an API, or bulk data downloads. This excludes the the largest database of research publications, Google Scholar, which offers a free public search interface, but does not provide an API or bulk downloads, and forbids automated crawling of the search service.

The following section (4.1.1) provide an overview of ten publication datasets. Table 4.1 provides an overview summary of the main features we were interested in. Namely, the table shows size, discipline coverage, ways of accessing the data (API, OAI-PMH, bulk downloads), and whether the dataset contains citations (column cit.) and full text (column FT). We summarise our findings in Section 4.1.2.

4.1.1 Datasets

ACL Anthology Network Corpus

The Association of Computational Linguistics (ACL) Anthology Network corpus[1] (AAN) is a collection of research publications in the fields of Natural Language Processing (NLP) and Computational Linguistics (CL) [Bird et al., 2008]. AAN is created from the ACL Anthology, which is a freely accessible repository of research publications in NLP and CL.

[1] `http://clair.eecs.umich.edu/aan/index.php`

Table 4.1: Overview of research publication datasets. The stars (*) in the table represent sources, which do not store full text but provide links to the full text of articles where available.

Source	Size	Domain coverage	API	OAI-PMH	bulk	cit.	FT
AAN	23k	NLP, CL	-	-	X	X	X
ArnetMiner	231m	general	X	-	-	X	-
ArXiv	1.3m	Phys., Math, CS	X	X	X	-	X
CiteSeerX	5.7m	CS	-	X	X	X	X
CORE	79m	general	X	X	X	X	X
DBLP	3.9m	CS	-	-	X	-	*
JSTOR	10m	general	-	-	X	X	*
Mendeley	N/A	general	X	-	-	-	*
MAG	120m	general	X	-	-	X	*
PubMed	27m	Biomed., life sci.	X	-	X	X	*

Because the corpus is composed of publications from two sub-fields of Computer Science, its size is significantly smaller than the size of other dataset, and was, at the time of writing this chapter, 23 thousand publications. The AAN corpus, which contains citation links between publications, as well as full-texts, can be downloaded in bulk[1]. [Radev et al., 2013] provided several overview statistics of the corpus with focus on the citation network, author collaboration network, and author citation network. AAN has been used for many tasks, including topic evolution studies [Hall et al., 2008], citation sentiment analysis [Athar and Teufel, 2012a], and for bibliometric studies [Radev et al., 2016].

ArnetMiner

ArnetMiner[2] is an index and a search engine for academic publications with focus on social network analysis [Tang et al., 2008]. It indexes

[2] `https://aminer.org/`

publications from the Web and identifies links between authors, conferences, and publications. The data can be accessed through an API, however, the API seems to be in development as some sample queries taken from the documentation did not work for us, and the documentation is incomplete. Furthermore, we were not able to determine whether the API enables retrieving citation links. According to the ArnetMiner homepage, at the time of writing this chapter the database contained over 231 million publications and 754 citation links, which makes it by far the largest database of scholarly publications in the world (possibly larger than Google Scholar, which is estimated to contain around 160-165 million publications [Orduña-Malea et al., 2015]). However, in 2010, it was been estimated the total number of journal articles published since the first journal was established was 50 million [Jinha, 2010]. Because journals are the most common way of publishing research for most disciplines, it is unlikely there are more conference publications in existence than journal publications. It is therefore unclear what types of article the figure shared by ArnetMiner include. Nevertheless, ArnetMiner has released several open datasets[3] which were used in a number of studies, especially studies concerned with social network analysis and ranking [Tang et al., 2009, 2012].

ArXiv

ArXiv[4] is an online self-archiving repository for research articles. It covers Physics, Mathematics, Computer Science (CS), Nonlinear Sciences, Quantitative Biology, Quantitative Finance, and Statistics, however, vast majority of publications submitted to ArXiv (around 95%) are from Physics, Mathematics or CS [ArXiv, 2017b]. The ArXiv data are avail-

[3] `https://aminer.org/data`

[4] `http://arxiv.org/`

able under various licenses (depending on the choice of the author), the most common one states that ArXiv is only permitted to distribute the articles but grants no additional rights [ArXiv, 2017a]. The data can be accessed through various method. ArXiv provides an OAI-PMH[5] endpoint and an API for accessing metadata of articles (which include a link to the article full-text), the PDF files can be downloaded in bulk [ArXiv, 2017a]. The size of the ArXiv dataset was almost 1.3 million at the time of writing this chapter. The dataset has been used in many different studies, including bibliometric-type works [Wang et al., 2013], and to study effects of Open Access publishing on publication visibility [Davis and Fromerth, 2007]. The dataset was also used in the 2003 KDD Cup which focused on citation and download prediction, and data cleaning [Gehrke et al., 2003].

CiteSeerX

CiteSeerX[6] [Giles et al., 1998] (previously CiteSeer) is a database of research publications, which focuses mainly on computer and information science. It crawls and harvests publicly available documents from the web and automatically extracts full text, metadata and citations from these documents. The size of the dataset was 5.7 million in 2016 [Wu et al., 2016]. CiteSeerX provides an OAI-PMH endpoint through which the CiteSeerX data can be harvested, as well as the possibility to download the data in bulk [CiteSeerX, 2017]. Several previous studies provided an analysis of CiteSeerX data with focus on the use of the data in scientometric and bibliometric studies [Fiala, 2011, 2012]. A recently study has also attempted to merge the dataset with the DBLP Computer Science

[5]The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol for harvesting publication metadata from online archives.

[6]`http://citeseerx.ist.psu.edu/`

Bibliography to produce a cleaner subset [Caragea et al., 2014].

CORE

CORE[7] (COnnecting REpositories) [Knoth and Zdrahal, 2012] is an aggregator of content stored in Open Access repositories. Besides harvesting and storing the content it provides additional services, such as a citation extraction and calculation of semantic similarity of publications. All CORE data are available under some Open Access compatible license. CORE data include publication full-texts (where available) in both PDF and text formats. The data can be accessed via an API, and through bulk download [CORE: Connecting Repositories, 2017]. At present, the CORE dataset contains nearly 79 million metadata records out of which more than 8 million records contain a PDF file. The CORE dataset has been used for in tasks such as to create word embeddings for citation classification [Lauscher et al., 2017]. Although it contains a citation network, due to relative sparsity of the network (it extracts references from publications for which it contains full-text) it has not been used in bibliometric studies.

DBLP Computer Science Bibliography

DBLP[8] (or The DBLP Computer Science Bibliography) [Ley, 2002] is an online bibliography of computer science research. It indexes metadata of books and documents from journals, conferences, etc. It does not store citation links between documents or full-texts, however the metadata contain links to the articles. The DBLP data are released under the ODC-BY 1.0 license, which means they can be freely used as long as their public use is attributed. The DBLP database can be accessed through an

[7] `https://core.ac.uk/`

[8] `http://dblp.uni-trier.de/`

API [Ley, 2009] or through a bulk data download. At the time of writing this chapter the bibliography contained 3.9 million records. Due to the lack of an extracted citation network, the DBLP dataset has not been used in many bibliometric studies, but has been used for other tasks, for example to study the community structure of computer science [Biryukov and Dong, 2010]. A detailed analysis of sub-fields contained in DBLP was presented in [Reitz and Hoffmann, 2010]. An enhanced version of the DBLP dataset containing abstracts and citations was released by ArnetMiner.org[9], however, at the time of writing this chapter, the latest version of the enhanced dataset was from 2010 and only contained 2 million papers (around half of the current size of DBLP).

JSTOR

JSTOR[10] is a multidisciplinary digital library which provides access to academic books and journals [Burns et al., 2009]. It covers many disciplines, predominantly humanities and social sciences. The JSTOR data is provided for non-commercial purposes through bulk data downloads. The data downloads can be requested via an online tool Data For Research (DFR) which allows querying the JSTOR corpus and defining the content to be downloaded [Burns et al., 2009]. Initially the downloads are limited to 1000 items per download, but larger downloads can be requested. The data downloads contain citations and additional information, such as key terms, however not full-texts. At present the size of JSTOR dataset is more than 10 million publications. As noted by [Bjork et al., 2014], the advantage of JSTOR is the timespan of the database. Unlike the large commercial databases, Web of Science and Scopus, the data in JSTOR goes back to the first issue of many existing journals.

[9] `http://arnetminer.org/dblp_citation`

[10] `http://www.jstor.org/`

JSTOR has been used in bibliometric studies [Bjork et al., 2014] as well as citation network analysis studies Shi et al. [2010].

Mendeley

Mendeley[11] [Henning and Reichelt, 2008], which is now owned by Elsevier, is a PDF reference manager for managing and sharing research papers. Mendeley is predominantly a desktop application; however, it also offers an API for querying its publication database. The API, among other things, enables retrieving metadata of documents contained in the Mendeley database. However, it does not offer a simple way of downloading the entire database. Mendeley is multi-disciplinary, and collects publication metadata from its users (new documents are added to Mendeley by the users of the desktop application) as well as from Elsevier's database Scopus. To the best of our knowledge, Mendeley does not publicly share information about the size of the database. Because Mendeley enables downloading information about a publication's readers, it has been used in a number of studies of altmetrics [Li and Thelwall, 2012, Maflahi and Thelwall, 2016].

Microsoft Academic Graph

Microsoft Academic Graph[12] (MAG) is a collection of research publications, authors, and other related entities, represented as a graph [Sinha et al., 2015]. It is the newest of the datasets presented in this section. MAG powers the academic search engine Microsoft Academic, which replaced the older Microsoft Academic Search. MAG was previously available for download in bulk, and the downloadable version of the dataset was used in research competitions, such as in the 2016 WSDM Cup on

[11] `https://www.mendeley.com/`

[12] `http://aka.ms/academicgraph`

raking academic papers [Wade et al., 2016], and the 2016 KDD Cup on predicting acceptance rate at conferences [Microsoft Research, 2016]. At the time of writing this chapter, the bulk download option was not available, however, Microsoft provides an API for accessing and querying the graph, which is free for a certain number of queries [Microsoft Azure, 2017]. MAG is multi-disciplinary, and at present contains more than 120 million publications[13]. Due to its size and broad coverage, MAG has already been used in a number of studies, including for topic detection and analysis [Effendy and Yap, 2017], and for citation prediction [Xiao et al., 2016].

PubMed

PubMed[14] is an index and a public search engine for scholarly literature in biomedical and life sciences. The database contains metadata of over 27 million publications, including more than 84 million citation links between articles. The data can be downloaded in bulk or queried through an API. Most articles in PubMed are subject to standard copyright and therefore are not available for download, however, Open Access articles can be downloaded both in bulk and through the API. PubMed has, due to its size, citation network, and good coverage of the biomedical field, been used in bibliometric studies [Xu et al., 2014, Lee et al., 2016], in literature-based discovery [Weeber et al., 2001, Srinivasan, 2004], and in other tasks.

[13]Since the time of writing this chapter a new study by [Hug and Brändle, 2017] was published which puts the MAG size at 168 million publications.

[14]`https://www.ncbi.nlm.nih.gov/pubmed/`

4.1.2 Summary

In this section we have briefly reviewed ten datasets of research publications which can be used in research analysis and evaluation studies. It can be seen these datasets vary greatly in size, coverage, and data quality. Several of the datasets offer good coverage, and good quality of data, but are limited to one or a few disciplines (AAN, ArXiv, DBLP, PubMed). On the other hand, the large datasets (ArnetMiner, CORE, MAG) are multi-disciplinary, albeit with some limitations (citation network sparsity in the case of CORE, data quality, which we were not able to verify, in the case of ArnetMiner). Microsoft Academic Graph is the newest of the datasets reviewed in this section, the first version of MAG was published in 2015. It is multi-disciplinary, and with more than 120 million publications also the second largest on the list. It contains a citation network, as well as venue, author, and field of study information. As such it seems to be a valuable resource for developing new research evaluation methods. However, because it was released only recently, it has not been used in many studies, and so it is not clear what coverage and data quality does it offer. To fill this gap, in the next section we provide an analysis of the dataset.

4.2 An Analysis of Microsoft Academic Graph

In the previous section we have shown although there are many datasets of research publications, all come with different limitations. A new dataset, called Microsoft Academic Graph[15] (MAG) Sinha et al. [2015] has been made openly available recently. MAG is a large heterogeneous graph comprised of more than 120 million publications and the related authors, venues, organizations, and fields of study. Up to date, MAG is

[15] `http://aka.ms/academicgraph`

one of, if not the largest publicly available dataset of scholarly publications, and of open citation data. However, as the dataset is assembled using automatic methods Sinha et al. [2015], before a decision can be made on whether to use it, for what purposes and with what limitations, it is important to understand how accurate it is and whether there is any noise or bias in the data. This section aims to answer this question. What interests us is the level of reliability of the data. The characteristics of the dataset are studied here by comparing the data with other publicly available research publication datasets. Among other things we are interested in topical and temporal coverage and in the properties of the citation network. This section is organised as follows. We start by describing the dataset and our methodology (Section 4.2.1). In Section 4.2.2 we presents the results of our study. Finally, in Section 4.2.3 we summarise our findings and conclude this section.

4.2.1 Dataset and method

The Microsoft Academic Graph is a large heterogeneous graph which models scholarly communication activities and which consists of six types of entities – publications, authors, institutions (affiliations), venues (journals and conferences), fields of study and events (specific conference instances); and the relations between these entities – citations, authorship, etc. The relations between the entities are described in more detail in Sinha et al. [2015]. The dataset contains publication metadata, such as year of publication, title and DOI. It does not contain the publication full texts or abstracts. For our study we have used the last version of MAG which was downloadable in bulk[16] (released on February 5, 2016).

[16]Between our study and putting together this chapter, Microsoft has removed the option of bulk downloads, stating that "the increased size and update frequency of the graph makes the blob download process impractical" [Microsoft Research, 2017]. We

Table 4.2 shows the size of the dataset.

Table 4.2: Microsoft Academic Graph size.

Papers	126,909,021
Authors	114,698,044
Institutions	19,843
Journals	23,404
Conferences	1,283
Conference instances	50,202
Fields of study	50,266

We are interested in analysing the dataset to understand its properties. Specifically, we are interested in answering the following questions:

- How sparse are the data (in terms of temporal properties, discipline coverage, institution/country representation, etc.)?
- How many of the entities have all associated metadata fields populated and how reliable are these data (for example publication years and fields of study)?
- How well are the data conflated/disambiguated (for example the author entities)?

Some of these questions can be answered by analysing the dataset directly. However, a manual evaluation or a comparison with another overlapping dataset could provide additional insights. As other publicly available sources of data are available, we have used these sources to study the accuracy and reliability of the dataset.

obtained the latest bulk download before Microsoft discontinued this way of accessing the data, and in our analysis we use this version. This enables us to quickly examine the entire graph, and gives us an idea what to expect when using the API.

Specifically, we have used the CORE[17] [Knoth and Zdrahal, 2012], Mendeley[18] [Henning and Reichelt, 2008], the Webometrics Ranking of World Universities[19] [Consejo Superior de Investigaciones Científicas, 2015] and the Scimago Journal & Country rank[20] [SCImago, 2007]. CORE is an aggregator of content stored in Open Access repositories and journals, its data include publication full texts (where available) in both PDF and text formats, as well as automatically extracted citations (for more details see Section 4.1). The version we used in our study is from April 2016 and contains over 25 million publication records. Mendeley is a crowdsourced collection of millions of research publications, offering metadata including abstracts, venue information, etc., however not citations and full-texts. As of writing this section the collection contains more than 100 million publications. The Webometrics Ranking of World Universities is an initiative publishing webometric rankings of universities, but also a list of top universities from around the world based on citation data assembled from Google Scholar. Finally, the Scimago Journal & Country Rank website publishes journal and country rankings which are prepared using data from Elsevier Scopus. We use the first two datasets to study how reliable are the metadata in the MAG, while the other two datasets are used to study the citation network.

All but the last of these datasets are, similarly as the MAG, assembled largely using automatic methods (crawling, harvesting, etc.), which means these datasets could as well suffer from bias or noise. For this reason, we are not aiming to find whether one of the datasets is better than the others, but rather to see whether there are similarities and the datasets are comparable. We believe in case we find a correlation and

[17] `http://core.ac.uk`

[18] `http://dev.mendeley.com`

[19] `http://www.webometrics.info`

[20] `http://www.scimagojr.com`

significant similarities between all of the datasets, this shows a certain level of accuracy and reliability.

4.2.2 Results

Publication age

The year of publication is one of the most important pieces of information about a publication for bibliometrics research. Consequently, it is critical that the data are reliable and consistent. For this reason our first task was to investigate the years of publication provided in the MAG.

The publication metadata contain titles, publication dates, DOIs and venue names (which are linked to venue entities). Impressively, the year of publication is populated for all papers in the dataset. Figure 4.1 shows a histogram of the publication years for documents published between 1900 and 2017. The oldest publication in the MAG was published in 1800, and there are 974,308 publications in the MAG which were published prior to the year 1900. Mean year of publications across all publications in the MAG is 1997.

To asses how reliable the publication dates in the MAG are, we have compared this data with dates obtained from CORE and Mendeley. To identify common publications between the three datasets, we have used the Digital Object Identifier (DOI). Table 4.3 lists the number of common documents we were able to identify. The last row in the table represents the number of documents after removing documents with any missing data, that is publications for which we were not able to obtain the publication date from one or more of the datasets.

We have compared the datasets using two methods – the Spearman's ρ correlation coefficients and the cumulative distribution function of the difference between the publication years in the different datasets. Table

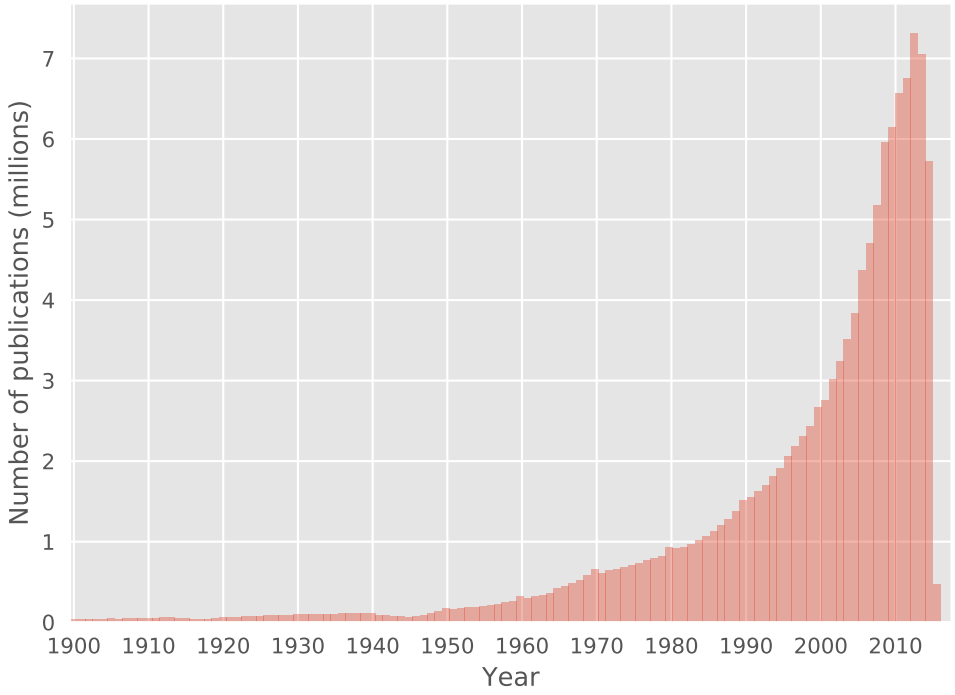

Figure 4.1: Histogram of years of publication provided in the MAG.

Table 4.3: Number of documents used for comparing publication dates in the MAG, CORE and Mendeley.

Unique DOIs in the MAG	35,569,305
Unique DOIs in CORE	2,673,592
Intersection MAG/CORE	1,690,668
Intersection MAG/CORE/Mendeley	1,314,854
Intersection Without missing data	1,258,611

4.4 shows the Spearman's ρ correlation coefficients. The Spearman's ρ correlations are all very strong (close to 1.0), the strongest correlation is between Mendeley and CORE ($\rho = 0.9743$), the weakest is between the MAG and CORE ($\rho = 0.9555$). To assess how big are the differences between the datasets we have calculated the cumulative distribution function of the differences between the three datasets.

To see in how many cases do the datasets agree, we have calculated

Table 4.4: Correlations between publication years found in the MAG, CORE and Mendeley. The p-value < 0.01 in all cases.

Spearman's rho	**MAG**	**CORE**	**Mendeley**
MAG	-	0.9555	0.9656
CORE	0.9555	-	0.9743
Mendeley	0.9656	0.9743	-

the cumulative distribution function of the difference between the data (Figure 4.2). To plot this function we use the absolute difference between the year of publication found in two datasets. Each point in the figure represents the proportion of publications for which the difference equals or is less than the value on the x-axis. The faster the line in the figure grows the more publications have the same or similar year of publication in the two datasets.

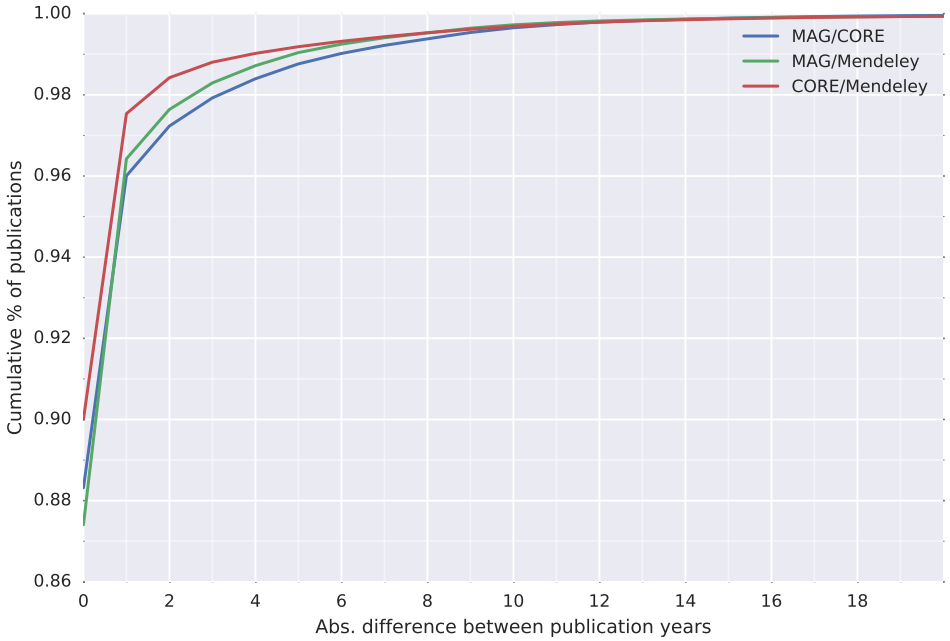

Figure 4.2: Cumulative distribution function of absolute difference between publication years found in the three datasets.

For all three comparisons the year of publication is the same in more than 87% of cases, which represents more than 1 million publications. The most similarities are found between CORE and Mendeley, where the year of publication differs by zero years in $\tilde{9}0\%$ of cases. A potential explanation for a difference of up to one year could be that one dataset contains the postprint version while the other a preprint, which was deposited online before the postprint version was published. MAG compares to the two other datasets very similarly, with $\tilde{8}8\%$ of papers having a difference of zero years and more than 96% of paper differing by zero or one year in both cases. That is, out of the 1.2 million publications less than 40 thousand have a difference of more than two years.

Authors and affiliations

The publications in the graph are linked to author and institution entities, which are both (to a certain level) disambiguated. Figure 4.3 shows mean number of authors per publication per year, and Table 4.5 presents summary statistics of the two networks.

Table 4.5: Summary statistics for the authorship and affiliation networks

Mean number of authors per paper	2.66
Max authors per paper	6,530
Mean number of papers per author	2.94
Max number of papers per author	153,915
Mean number of collaborators	116.93
Max number of collaborators	3,661,912
Number of papers with affiliation	20,928,914
Mean number of affiliations per paper	0.23
Max number of affiliations per paper	181

It is interesting to notice all publications in the graph are linked to one or more author entities, however 105,980,107 publications are not affili-

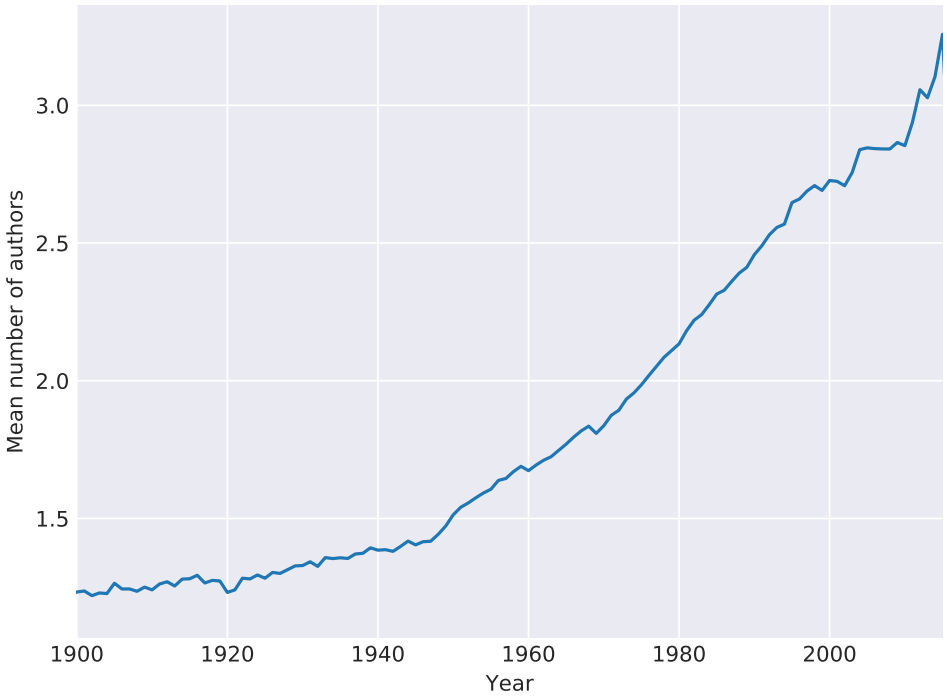

Figure 4.3: Mean number of authors per publication and year.

ated with any institution. Furthermore, while the mean values presented in Table 4.5 are similar to numbers reported for other datasets [Newman, 2004], the maximum values point to some discrepancies in the data. For example, the highest number of authors on a publication was reported to be 5,154 [Castelvecchi, 2015]. In MAG the same article comes fourth in terms of number of authors after papers titled "Sunday, 26 August 2012", "Monday, 27 August 2012" and "Tuesday, 28 August 2012". Furthermore, the author with most publications is "united vertical media gmbh". However, because the graph is built using automatic methods, such errors are expected. In order to understand how reliable the data in MAG are, we have compared the most cited institutions in MAG to the most cited institutions according to the Ranking Web of Universities website [Consejo Superior de Investigaciones Científicas, 2015], which uses data from Google Scholar. The results of this comparison are presen-

ted later in this section. We have not done the same comparison for the author entities due to potential disambiguation issues.

Journals and conferences

Similarly as with the author and affiliation entities, the papers in MAG are linked to publication venues – journals and conferences. Aside of a list of conferences consisting of a name and abbreviation (e.g. "JCDL – ACM/IEEE Joint Conference on Digital Libraries") the MAG also contains a list of conference instances containing information when and where the conference took place. There are 51,900,106 publications in MAG which are linked to a journal entity and 1,716,211 publications linked to a conference. Interestingly 103,131 publications are linked to both a journal and a conference. We have manually investigated several of these publications and found that in cases this was due to a paper being presented at a conference and later in proceedings published as a journal. It is also interesting to notice that the number of journal publications in MAG is very close to the total number of journal publications estimated to be in existence [Jinha, 2010]. Similarly as with affiliations, we have compared journal citation data from MAG with citation data obtained from the Scimago Journal & Country Rank website SCImago [2007], which uses Elsevier Scopus data. The results of this comparison are presented later in this section.

Fields of study

Information about which field, or fields, of study a publication belong to is very valuable for many tasks. At the same time this information is often complicated to get as it is dependent on either having access to the text of the publication or access to manually created metadata. We investigate the fields of study provided by MAG for papers in the graph

in order to understand what is the coverage of the dataset. The fields of study found in MAG are organised hierarchically into four levels (level 0 to level 3, where level 3 has the highest granularity). There are 47,989 fields of study at level 3 (for example "concerted evolution"), 1,966 at level 2 (e.g. "evolutionary developmental biology"), 293 at level 1 (e.g. "genetics") and 18 at level 0 (e.g. "biology"). 41,739,531 out of the 126,909,021 papers in total (that is about $\tilde{3}3\%$) are linked to one or more field of study entities. Figure 4.4 shows the distribution of papers over the 18 level 0 fields of study. In case the publication was linked to more than one level 0 field of study, we have counted it towards each linked field of study.

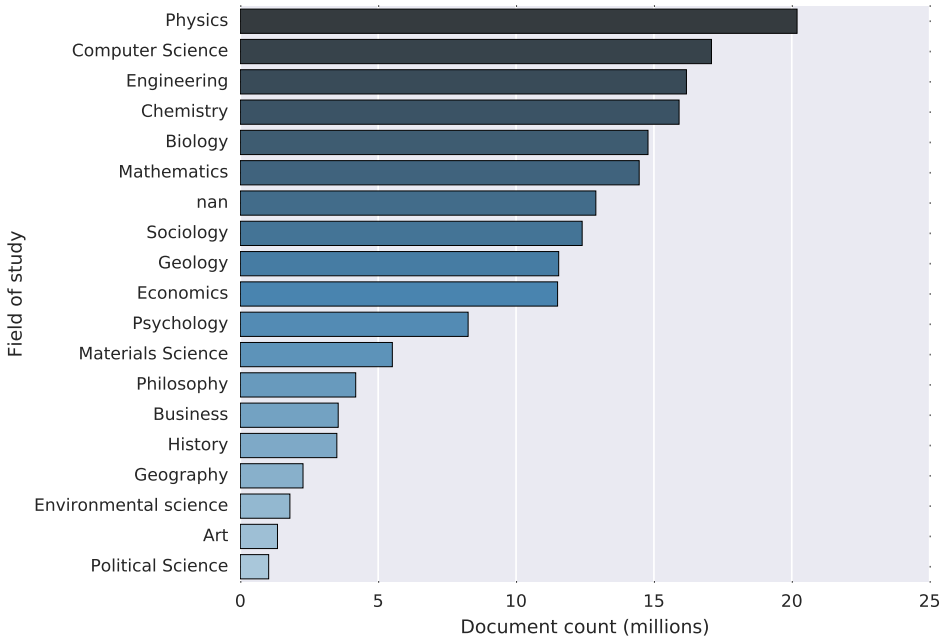

Figure 4.4: Distribution of papers into fields of study in MAG.

It can be seen that the three largest fields of study in MAG are Physics, Computer Science and Engineering, followed by Chemistry and Biology. This is to a certain degree consistent with other studies, which have reported Physics and Engineering to be among the largest discip-

lines in terms of number of publication, however Medicine and Biology are typically reported to be the most productive [Althouse et al., 2009, D'Angelo and Abramo, 2015]. One possible explanation for this bias towards the three technical fields could be due how the data is being collected. According to [Sinha et al., 2015] this is done, aside of using publisher feeds, by crawling the web. This could create bias towards scientific disciplines which tend to publish and deposit their publications online more frequently and therefore make their publications more easily discoverable. For comparison we have obtained information about readers from Mendeley for the 1,258,611 publications used in comparing the publication years in MAG, CORE and Mendeley. Our assumption is that the readers will bookmark publications related to their research area, based on this assumption we use the readers' research area to assign the papers to scientific disciplines. We use the proportion of readers in given area to assign the publication to the area, for example if a publication has 15 readers in Biology and 5 readers in Chemistry, we would add 0.75 to the first area and 0.25 to the second. At the lowest level of granularity Mendeley classifies publications into 22 disciplines, the distribution of the 1,258,611 papers into the 22 disciplines can be seen in Figure 4.5.

Citation network

One part of the dataset which is very interesting to us is the citation network. In order to understand how reliable the citation data in the MAG are, we study the citation network from several perspectives. First, we study the network by itself by looking at the citation distribution, to see whether it is consistent with previous studies. We then compare the citations received by two types of entities (institutions and journals) in the graph with citations from external datasets.

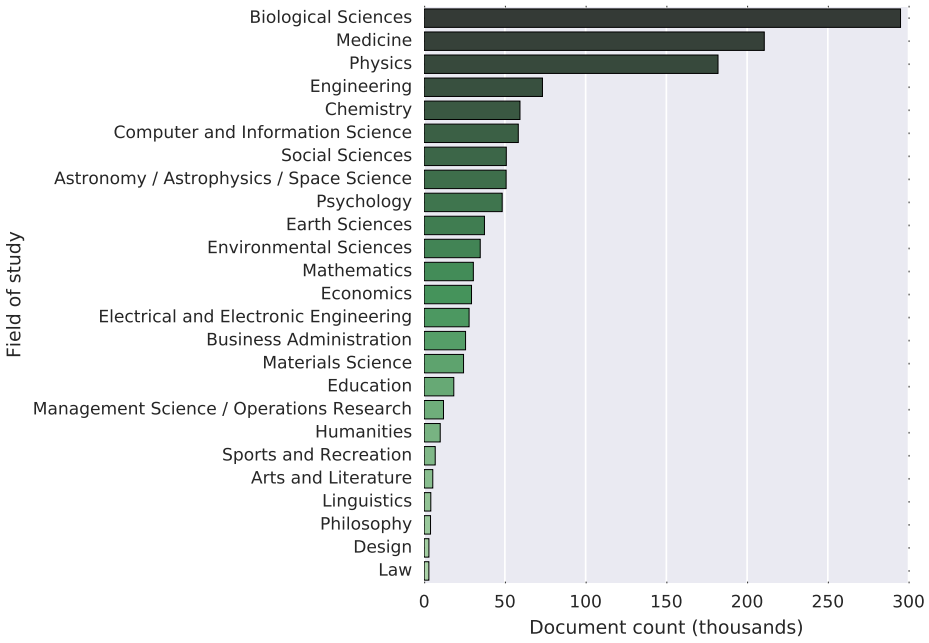

Figure 4.5: Distribution of papers into fields of study in Mendeley.

The MAG contains 528,682,289 internal citations (citations between the papers in the graph). This means each paper in the graph is cited on average 4.17 times. However, a significant portion of the papers are disconnected from the network (neither cite nor are cited by any other papers). Table 4.6 shows the number of disconnected nodes, there are over 80 million such nodes.

Table 4.6: MAG citation network statistics.

Total number of papers	126,909,021
Papers with zero references	96,850,699
Papers with zero citations	89,647,949
Papers with zero references and citations	80,166,717
Mean citation per paper	4.17
Mean citation per “connected” paper	11.31

It is not uncommon for research publications to never receive any

citations [Seglen, 1992]. In fact some studies estimate the proportion of publications which are never cited to be between 23% and 90% depending on the discipline [Weale et al., 2004, Bauerlein et al., 2010, Meho, 2007]. Although it is possible for a research publication to not contain any references, we believe the proportion of such publications will be minimal, however we were not able to find any study estimating what is the proportion of such publications. Furthermore the approximate number of received citations per publication across all disciplines has been reported to be ≈ 11 [Times Higher Education, 2011].

These statistics show that although when we exclude the "disconnected" publications, the citation network is reasonably dense, the proportion of papers which do not have have any outgoing edges in the network (references) is quite staggering. Furthermore, it is interesting to notice how has the citation network provided in the MAG been changing with each new version of the dataset. Microsoft has so far released four versions of the dataset (in May 2015, August 2015, November 2015 and February 2016). We have investigated the tree latest versions. While the number of paper entities in the graph has remained about constant (with a growth from 122 million papers in August 2015 to 126 million papers in February 2016), the size of the citation network has been changing significantly – it has first grown from over 750 million edges to over 950 million edges, but has in the latest version been reduced to 528 million edges. While this shows Microsoft keeps constantly improving the dataset, these changes could also suggest potentially unreliable data. In order to further study the properties of the citation network, we have compared the citation data found in MAG with the Ranking Web of Universities [Consejo Superior de Investigaciones Científicas, 2015] and the Scimago Journal & Country Rank (SJR) [SCImago, 2007] citation data. The Ranking Web of Universities website aggregates institutional

profiles found in Google Scholar to count the total citations received by a university. The website provides a list of top 2105 universities around the world along with the aggregated citation counts. We have used a version of the list published in December 2015. The Scimago website publishes journal ranks and total citation counts based on data obtained from Elsevier Scopus. The number of journals listed on the website is 22,878. The citation totals found on the Scimago website represent the sum of citations received by papers published in the journal over a three-year period. Specifically it is citations received in 2014 by the journal's papers published in 2011, 2012 and 2013.

We have compared the MAG citation data with the external lists using two methods which complement each other. The methods are the size of the overlap of the lists and the Pearson's and Spearman's correlation coefficients. The overlap method ignores the ranks and counts how many items appear in both lists. The correlations are calculated only on the matching items.

To identify the common items in the MAG and the two external lists, we have normalised the university and journal names (we removed all accents, special characters etc.) and tried to match the normalised names. We accept the names as matching only in case we identify a full string match. This way we were able to match 1,255 universities (out of 2,105 found on the Ranking Web of Universities website) and 13,050 journals (out of 22,878 found on the SJR website). We then count total citations received by the university/journal in MAG (in case of journals we limit the citations to the same time period as in the SJR data). Finally we rank both lists (the MAG and the external data) and calculate the absolute difference between the ranks for each university and journal. Figure 4.6 shows a scatter plot of the university citation counts and Figure 4.7 a scatter plot of the journal citation counts.

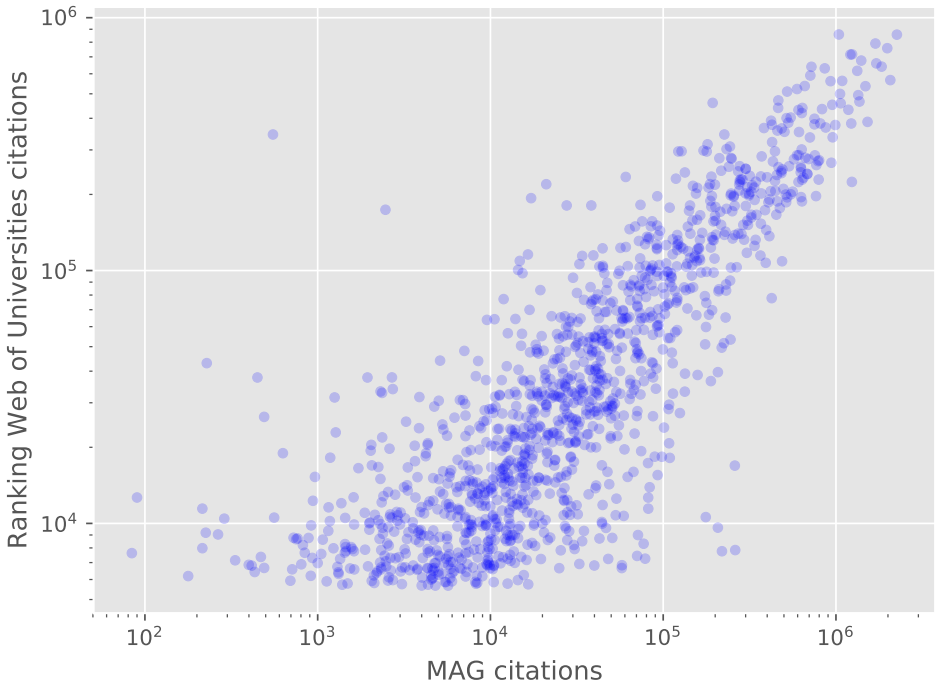

Figure 4.6: Comparison of university citations in MAG and on the Ranking Web of Universities website.

We first compare the lists using overlap. When comparing the journal lists, we found 4 common journals among the top 10, 54 among the top 100, 677 among the top 1000 and 1407 among the top 2000. Table 4.7 shows the top 10 journals in both lists, with the journals appearing in both lists highlighted in bold. For this comparison, we use all journals found in the MAG and on the SJR website, not only the common items.

Table 4.7: Top 10 journals according to the MAG and the Scimago Journal & Country Rank website. Highlighted in bold are those journals, which appear in both lists.

Rank	MAG	SJR
1	**Plos One**	**Plos One**
2	Proceedings of the National Academy of Sciences	Journal of the American Chemical Society

Rank	MAG	SJR
3	**Nature**	**Nature**
4	**Science**	**Science**
5	Journal of Nanoparticle Research	Physical Review Letters
6	**Journal of Biological Chemistry**	Chemical Communications
7	Nanoscale Research Letters	**Journal of Biological Chemistry**
8	The New England Journal of Medicine	Journal of Physical Chemistry C
9	BMC Public Health	Applied Physics Letters
10	Cell	Journal of Materials Chemistry

Unfortunately, we were not able to produce a similar statistic for the universities lists, as in the MAG universities are mixed with other affiliations (research institutes, companies, etc.) in one table. For comparison we have manually picked the first 10 universities according to their total citation counts found in the MAG and compared this list to the top 10 universities (in terms of total citation counts) according to the Ranking Web of Universities website. The two lists are shows in Table 4.8.

Table 4.8: Top 10 universities according to the MAG and the Ranking Web of Universities website. Highlighted in bold are those universities, which appear in both lists.

Rank	MAG	Ranking Web of Universities
1	**Stanford University**	Harvard University
2	University of Washington	University of Chicago
3	**Massachusetts Institute of Technology**	**Stanford University**
4	**University of Michigan**	**University of California Berkeley**
5	**Johns Hopkins University**	**Massachusetts Institute of Technology**

Rank	MAG	Ranking Web of Universities
6	**University of California Berkeley**	University of Oxford
7	University of California	University College London
8	University of Texas at Austin	University of Cambridge
9	University of Wisconsin Madison	**Johns Hopkins University**
10	University of Toronto	**University of Michigan**

There is one surprising difference in Table 4.8, which is the lack of Harvard University in the top 10 universities according to the MAG (in the MAG, Harvard is in 14th position), as Harvard University is known to be among the most, if not the most cited university. However, this is due to the fact different Harvard schools appear in the MAG separately (for example "Harvard Law School" or "Harvard Medical School" are listed as separate affiliations). We have manually summed all Harvard schools present in the MAG which moved Harvard University to the top of the list. The other differences, particularly the different positions of universities in the two lists are to be expected, as differences between different citation databases are know to exist. For example, when comparing the Table 4.8 to the list of top universities provided by the Science Watch website [Science Watch, 2009a], it can be seen no two lists overlap exactly. The University of Washington, which is second in terms of total citations according to the MAG appears fourth in the Science Watch list but does not appear in the top 10 list according to the Ranking Web of Universities at all. This situation is similar for the journals [Science Watch, 2009b].

To quantify how much do the lists differ, we created histograms of the differences between the ranks in the MAG and in the external lists, which are shown in Figures 4.8 and 4.9. Figure 4.8 shows differences between the ranks of universities, while Figure 4.9 show differences between the

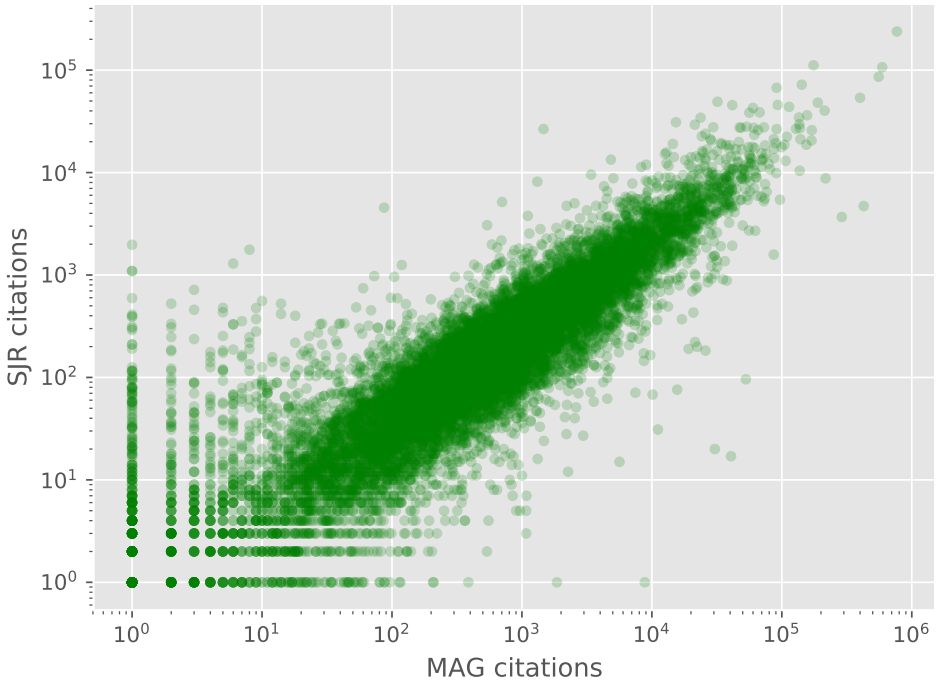

Figure 4.7: Comparison of journal citations in MAG and on the SJR website.

journal ranks. To produce these histograms, we first sorted the data by the total citations provided in the external list (the Ranking Web of Universities and the SJR list). We then took the top 100/top 1000 universities/journals and created a histogram indicating how much their ranks differ from the ranks provided by the MAG.

The results show that university citation ranks in the MAG differs by more than 200 positions for about 20% of universities in the top 100 of the Ranking Web of Universities list. The citation university rank differs by less than 25 positions for less than 40% of universities across these two datasets. A similar situation is observed with journal ranks. This high discrepancy in rankings is not necessarily the problem of the MAG, but possibly of the reference lists, as these show lower absolute citation counts than in the MAG. As it is possible to investigate the data

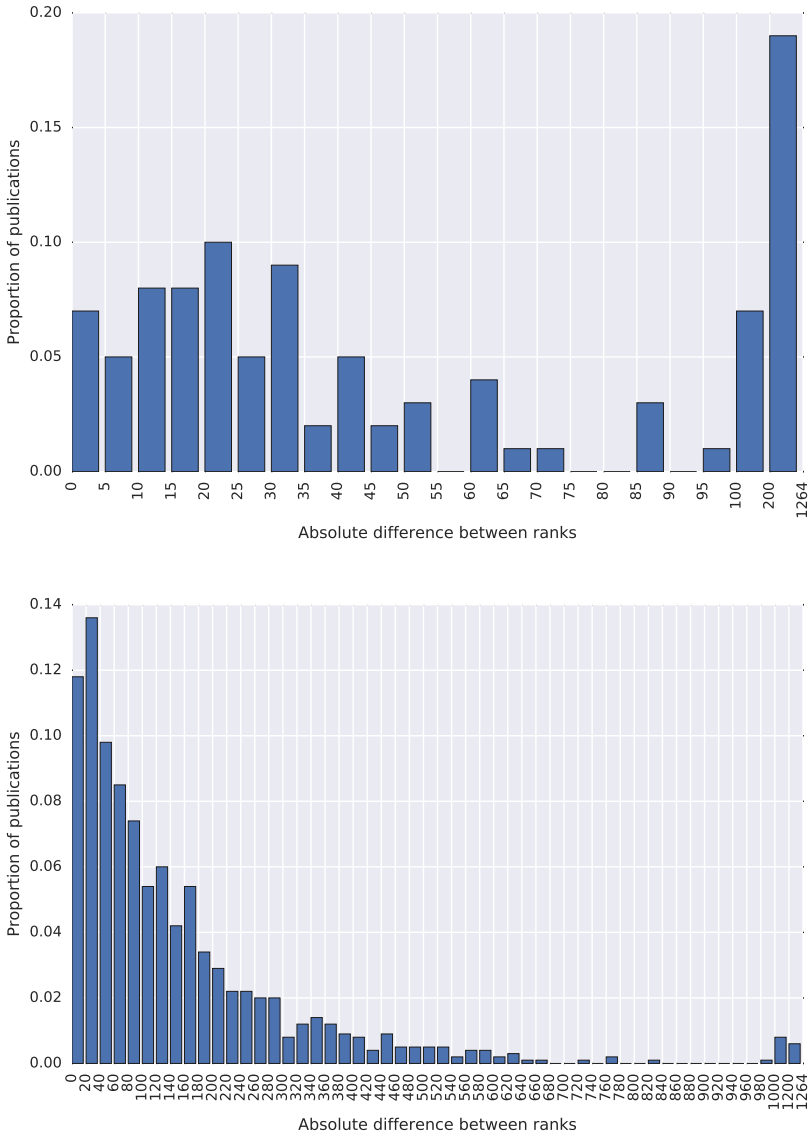

Figure 4.8: Top 100 (top) and top 1000 (bottom) universities according to the Ranking Web of Universities website, and the difference between their rank in the MAG and according to the website.

at the granularity of individual citations in the MAG, which is not the case for the external lists we used, we believe that the MAG should be considered a more trustworthy source of data. The large differences in rankings produced by different providers indicate that a more transparent

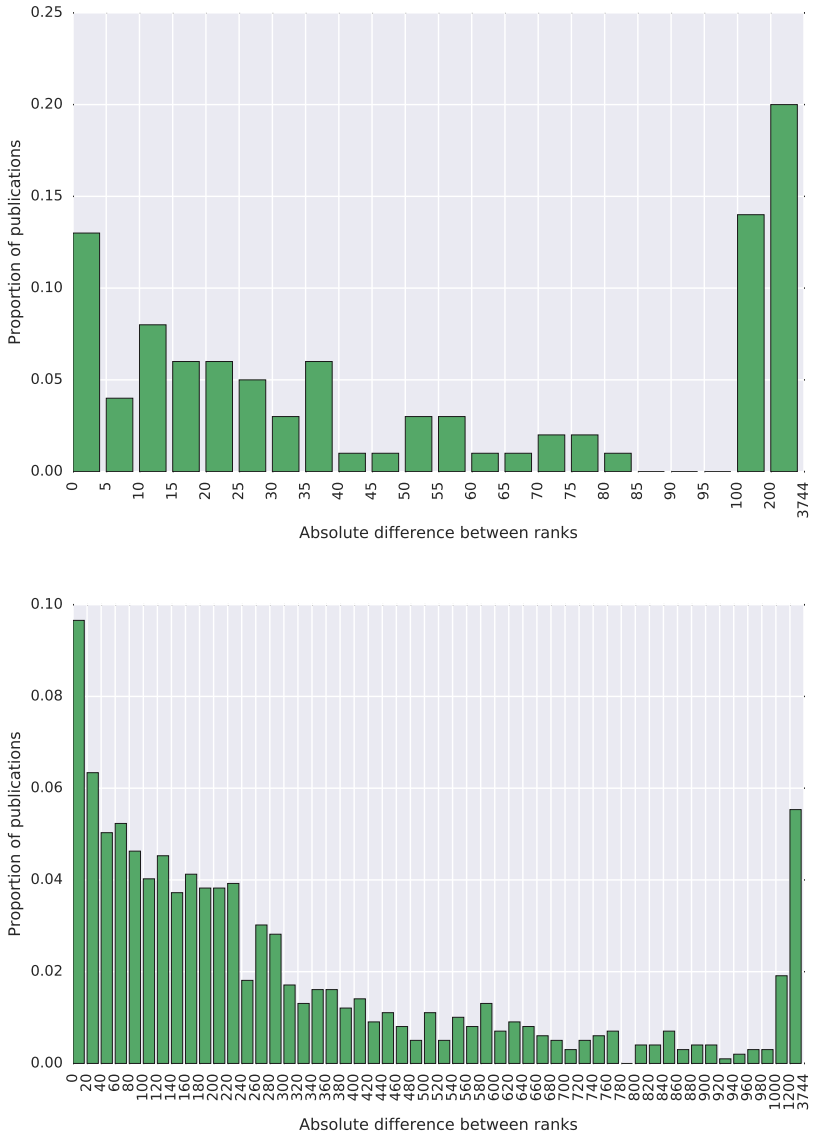

Figure 4.9: Top 100 (top) and top 1000 (bottom) universities according to the Scimago Journal & Country Rank website, and the difference between their rank in the MAG and according to the website.

approach to releasing citation data, so that errors can be investigated and corrected, is necessary to establish authority.

The correlations are reported in Table 4.9. These were calculated only on the matching items. We have found that on average the ranks

in the two lists of top universities (number of items in the two lists is $n = 1,255$) differ by 163, with standard deviation of 185. The Sperman's ρ correlation for the universities lists is $r = 0.8266$ ($p < 0.01$), which is a strong correlation. In case of journals ($n = 13,050$), the ranks differ on average by 1,203 with standard deviation of 1,211. The journals lists also correlate strongly, with Spearman's $\rho = 0.8973$ ($p < 0.01$). These strong correlations confirm that although there are differences between the datasets, these are, especially on the aggregate level, not significant, and the MAG can be used as a reliable source of citation data.

Table 4.9: Correlations between the MAG and the top universities list obtained from Ranking Web of Universities website and the journals list obtained from the SJR website.

	Universities	**Journals**
Pearson's r	$0.8773, p < 0.01$	$0.8246, p < 0.01$
Spearman's ρ	$0.8266, p < 0.01$	$0.8973, p < 0.01$

4.2.3 Summary

In this section we investigated the Microsoft Academic Graph, which is a large heterogeneous graph comprised of over 120 million publications, the related authors, institutions, venues and fields of study and relations between these entities. We reported on the analysis of the MAG comparing it with other research publication and citation datasets. While the MAG data correlate well with external datasets and are a great resource for doing research in scholarly communication, we have identified certain limitations as to the completeness of links from publications to other entities. Despite this, the MAG is currently the most comprehensive publicly available dataset of its kind and represents an astonishing effort

which will prove useful in many areas of research where full text access to publications is not required. The MAG is also an important step in the right direction in terms of releasing free and open citation data for research evaluation purposes, a recommendation made by a number of recent initiatives and reports, including the Metric Tide Report (Chapter 2). We showed that existing university and journal rankings, which are typically based on proprietary aggregated data, produce substantially different results. This diminishes the trust in these rankings. As the MAG is open and transparent at the level of individual citations, it is possible to verify and better interpret the citation data. We believe our analysis will be valuable to those deciding whether to use the MAG, for what purposes, how to avoid pitfalls, and how to interpret the results. This analysis is also beneficial to us, as we use the dataset throughout this thesis (Chapters 5 and 7).

4.3 Methods for evaluating research publication metrics

In the previous two sections, we have reviewed some of the best known research publication datasets, which are available for studying new research evaluation metrics. Now that we have an understanding of where to collect publications for testing new metrics, we focus on methods for evaluating the metrics. Specifically, we are interested in studying what methods can and are typically be used to determine whether, and how well, a certain metric works.

In general, an evaluation can be *qualitative* (Section 4.3.1) or *quantitative* (Section 4.3.2). In evaluation of research metrics, a qualitative evaluation would typically translate to calculating the metric of interest on a sample dataset, and then manually examining the results. On the

other hand, a quantitative evaluation is typically done by comparing the results with another metric or indicator, such as with peer review. Quantitative evaluation methods are more common, as such evaluations can be done on large amounts of data, and require less effort in analysing results. We also review a number of related works, particularly works focused on studying the meaning and function of citations (Section 4.3.3). We summarise our findings in Section 4.3.4.

4.3.1 Qualitative methods

A qualitative evaluation was performed by Hirsch in his seminal paper which introduced the h-index [Hirsch, 2005]. In his paper, Hirsch has calculated the h-index on five groups of scientists from two different disciplines, and of different seniority (however, very prominent scientists in all cases). For each group of scientists, Hirsch has presented either several people with the highest h-index, or descriptive statistics including mean and standard deviation, and provided a discussion of the values. A similar approach was used by Oberesch and Groppe, who have proposed a new index for evaluating scientists, the *mf-index* [Oberesch and Groppe, 2017]. The authors have calculated and studied the proposed index using data of six scientists from different fields, and of different seniority and prominence. Both [Hirsch, 2005] and [Oberesch and Groppe, 2017] have then provided a discussion and an analysis of the results. The strength of qualitative approaches to evaluation of research metrics lies in the ability to provide a strong explanation of results in context, however, the downside is the effort required to perform the analysis, which consequently limits the number of data points which can be analysed.

4.3.2 Quantitative methods

Probably the most common approach to quantitative evaluation of research metrics is a comparison with another existing metric or metrics, which is typically done through correlation analysis. Peer reviews [Rinia et al., 1998, Aksnes and Taxt, 2004], expert judgements and rankings [Waltman and Costas, 2014, Wade et al., 2016], Journal Impact Factor [González-Pereira et al., 2010], citation counts [Bornmann and Daniel, 2006, Costas et al., 2015], and other metrics, have been previously used in these evaluations. Rinia et al. [1998] have compared peer review results (which consisted of a set of criteria, each ranked on a scale from 1 (best) to 9 (worst)) of condense matter physics programmes in Netherlands, with several bibliometric indicators (e.g. number of citations with and without self-citations, number of publications, journal average citation rate) and found significant correlation for several of the indicators, including total number of citations. A similar study with similar results has been conducted by Aksnes and Taxt [2004]. Waltman and Costas [2014] have used recommendations from F1000 (Faculty of 1000), which is a platform for biomedical and life sciences publishing recommendations of articles provided by F1000 members (experts in the field, the "faculty"). The authors have observed a weak correlation between number of recommendations and citation counts. Expert judgements were also used as ground truth in the 2016 WSDM Cup Challenge [Wade et al., 2016]. González-Pereira et al. [2010] have used Journal Impact Factor to analyse the performance of their new metrics, the SCImago Journal Rank (SJR). In contrast to Rinia et al. [1998], Bornmann and Daniel [2006] have used citation counts to evaluate the effectiveness of peer review for awarding fellowships to post-doctoral researchers (instead of using peer review to evaluate citation counts). Because prior articles of the accepted applicants are more likely to be highly cited than those

of the rejected applicants, the authors concluded that peer review works for selecting the best junior scientists. However, the same researchers have used the argument that citation counts correlate with peer reviews to conclude that citation counts capture publication quality [Bornmann and Haunschild, 2017]. Citation counts have also been used to analyse different altmetrics [Costas et al., 2015].

It can be seen that evaluating metrics using a comparative analysis is fairly common. However, this approach has both advantages and limitations. The typical reason for using such approach is the ability to provide an analysis on a large amount of data. Furthermore, the metrics used in the comparison mentioned in the previous paragraph are typically widely used and well known in the scientific community, which makes the analysis and conveying the results easier. However, each of the metrics used in these evaluations comes with certain limitations which need to be taken into account. For example, it is not clear how much are the human judgement data used by [Rinia et al., 1998], [Wade et al., 2016] and [Bornmann and Daniel, 2006] biased towards citation counts. This issue could manifest in case the judges had access to such information when rating the publications. Furthermore, although the metrics used above provide simple and easily understandable comparisons, there is an ongoing research and discussion trying to answer whether these metrics themselves capture scientific impact and quality [Seglen, 1997, Bornmann and Daniel, 2008, Campanario and Acedo, 2007, Campbell, 2008, San Francisco DORA, 2012, Francois, 2015, Ricker, 2017], which makes their use somewhat unsubstantiated. Finally, a significant obstacle is the difficulty of obtaining certain data, particularly expert judgements (unfortunately the only large dataset of peer review judgements known to us – F1000 recommendations [Waltman and Costas, 2014] – is not openly available). Despite these limitations, comparative analysis is a frequently

used method for evaluating research metrics, and is considered a standard method in certain fields such as webometrics and altmetrics [Thelwall and Kousha, 2015a].

4.3.3 Other approaches

One strand of research has focused on analysing the underlying data used in research metrics, specifically citations, in order to understand what does this data capture, and consequently, whether meaningful research metrics can be built using this data. The approaches focused on studying the validity of citations for research evaluation can broadly be categorized into two groups. One group has focused on the unit of measurement itself, and has studied, for instance, the reasons for citing [Harwood, 2008] or not citing [MacRoberts and MacRoberts, 2010] specific papers, or the characteristics of citation, such as the placement [Bertin et al., 2016a], and the context [Hu et al., 2015] of citations in text. The second group has concentrated on understanding what citations represent, for example by studying the characteristics of highly cited publications [Wang et al., 2011, Antonakis et al., 2014] and which other factors do they correlate with [Bornmann and Leydesdorff, 2015] (e.g. Journal Impact Factor, number of authors, and paper length). While these approaches have helped understanding and explaining different characteristics of citations, they have so far failed to conclusively demonstrate whether citations work as an indicator of research quality or impact.

4.3.4 Summary

In this section, we have shown many different approaches to evaluating research metrics exist. Typically, these involve a manual examination and explanation of results, or a comparison with another metric or met-

rics. Each approach comes with certain advantages and limitations, and neither helps to answer the question completely. We believe the main reasons for this lack of a reliable evaluation method is a lack of a true ground truth dataset – a dataset containing a widely agreed and accepted publication evaluations/rankings, which could be used to compare new metrics to. Although the F1000 dataset (a dataset of peer rankings in the field of biomedical and life sciences) comes close, it is not publicly available, and cannot therefore be easily used for developing and testing new metrics. For this reason researchers resort to using other methods, particularly the methods mentioned above. We believe creating a ground truth or a validation dataset would be a valuable addition to the state-of-the-art in this area which would facilitate the development of new research metrics.

4.4 Development of a new dataset for evaluating research metrics

In the previous section we have discussed methods that are typically used to evaluate new research metrics. We have described the existing methods and shown that in this area, no ground truth validation dataset exists that can be used to analyse and evaluate new research metrics. Due to the lack of a validation dataset, the authority of new research metrics is often established axiomatically, or with little evidence that they measure what they intend to measure. For example, the two best-known metrics, the Journal Impact Factor (JIF) [Garfield, 1972] and the h-index [Hirsch, 2005], were both proposed without such evidence. Furthermore, the unavailability of a validation dataset complicates the development of new metrics. For this reason, in this section we focus our attention at this problem and describe a new dataset we developed that

can be used for validating new research metrics and that complements the existing data and approaches. This section is organised as follows. In Section 4.4.1 we explain the idea behind the creation of the dataset. Next, in Section 4.4.2 we explain our motivation for creating the dataset and describe our research methodology. In Section 4.4.3 we explain how the dataset was created. Finally, Section 4.4.4 presents some overview statistics of the dataset.

4.4.1 Introduction

As we have shown in Chapter 3, when talking about research evaluation and scientific impact and excellence, most people usually refer to the volume of change produced in a particular field (research contribution, how much did a piece of work move the field forward), rather than referring to the educational (or other types of) impact generated. This is also the case for many national evaluation systems [Research Excellence Framework, 2012, Tertiary Education Commission, 2013b, Australian Research Council, 2015a]. A characteristic example of the first type (publications which produced a high volume of change) are *seminal publications*, while *literature reviews* (surveys) are a typical example of the second type (publications generating different types of impact). Indeed, the definition of the word *seminal* according to the Oxford Dictionary is "strongly influencing later developments" while the definition of the word *review* is "a report on or evaluation of a subject or past events", which matches our understanding of the difference between these two types of papers. Hence, if one of the goals of research evaluation is recognising publications which provided a significant research contribution to their field, seminal papers should on average perform better under such evaluation than literature reviews, which by definition do not generate a significant

change in the field[21].

In this section, we describe the creation of a new dataset of seminal publications and literature reviews which we call TrueImpactDataset. This dataset was built from data collected in an online survey. We asked the respondents to provide two references from their research area – a seminal publication and a literature review. We have shared this dataset with the research community[22] to help the development of new research evaluation metrics. The dataset consists of metadata (which include DOIs) of 314 research papers from different scientific disciplines – 148 survey papers and 166 seminal papers. Furthermore, in the final part of this section we discuss the parameters an ideal dataset for developing novel metrics should satisfy.

4.4.2 Methodology

As we have explained in the introduction to this chapter, this chapter aims at answering the following research question: "How can we evaluate the performance of metrics used in research evaluation for assessing the value of research publications?" In the previous section we have reviewed different approaches which have been used in the past for evaluating research metrics. We propose to use a slightly different method. A typical data analysis/statistics approach to answering the question above would be to test the metric on a ground truth dataset, such as a ranked set of papers, and to express the success rate of the metric using an evaluation measure such as precision and recall. However, to our knowledge, there exists no openly available ground truth or a reference dataset that could be used for establishing the validity of research metrics. While there was

[21]With some exceptions, notably systematic reviews, which are a key practice in evidence-based medicine

[22]`http://trueimpactdataset.semantometrics.org/`

an attempt at creating such a dataset [Wade et al., 2016], this dataset was not openly shared and so cannot assist with this task. A similar dataset which has recently been used for this purpose also is not openly available [Waltman and Costas, 2014]. Because building such dataset would require significant time and resources (Section 4.4.5) we were looking for an alternative approach.

As mentioned in the previous section (4.4.1) when talking about evaluation of research outputs, an important dimension is the amount of change produced in a research area (how much was the area pushed forward thanks to a given piece of work) [Research Excellence Framework, 2012, Tertiary Education Commission, 2013b, Australian Research Council, 2015a]. This amount of change has been discussed and studied from different perspectives [Yan et al., 2012, Whalen et al., 2015, Valenzuela et al., 2015, Patton et al., 2016]. We were looking for a sample of research publications representing such work and we believe seminal research papers constitute such sample. To provide a clear comparison we were also interested in review publications (papers presenting a survey of a research area). While these papers are often highly cited [Seglen, 1997, Aksnes, 2003] they often do not present new original ideas. Our goal is to study whether new and existing research metrics distinguish between these two types of papers.

To our knowledge, there currently is not any dataset which would categorize papers into these two categories. We were therefore left with creating such dataset ourselves. We have employed an online survey for this task. The format of the survey, the number of collected responses and other details are presented in Section 4.4.3. In the following section (4.4.4) we analyse the dataset to understand whether it is suitable for our purposes.

4.4.3 Dataset creation

This subsection describes the dataset and the process used to create it. The dataset is publicly available for download[23].

Initial data collection

The goal was to create a collection of research publications consisting of two types of papers, seminal works, and literature reviews. We have used an online form to collect the references, which was composed of two sets of questions – questions about the respondent's academic background (their discipline, seniority and publication record) and questions which asked for a reference to a seminal paper and to a literature review, both related to the respondent's discipline. We have used the latest Research Excellence Framework (REF) units of assessment [Research Excellence Framework, 2014a] as a list of disciplines when asking about the respondents' academic background because UK researchers are familiar with this classification. The complete survey together with the invitation email can be seen in Appendix B.

The survey was sent to academic staff and research students from all faculties of the Open University (to 1,415 people in total). The reason why we contacted Open University researchers is because research at the Open University covers many disciplines, and because it is the largest university in the United Kingdom. We were therefore able to get a significant sample spanning multiple disciplines. Within three months we have received 184 responses (172 references to seminal papers and 157 to review papers), which represents a 13% response rate. The survey questions and email invitation are available online together with the dataset[23]. To enable the respondents to send at least one reference, in case they were

[23] `http://trueimpactdataset.semantometrics.org`

not able to submit both, we made both answers optional. Ten respondents have only filled the questions related to their academic background but have not provided the references. We have removed these responses from the dataset which left us with 174 responses.

We did not require the references to be in a specific format (e.g. a URL or DOI) to make it easier to complete the survey. The respondents were allowed to submit the references in any format they preferred (as a text, link, etc.). As a consequence, a few of the references were submitted in a format which made it impossible for us to identify the papers (e.g. "Stockhammer (2004)"). We have removed these papers from the dataset. After removing empty and unidentifiable responses, we were left with 171 responses providing us with 166 seminal and 148 literature reviews.

Additional metadata

Once the survey was closed we have manually processed the data and collected the following information (by querying a search engine for the paper title and looking for a relevant page): a DOI, or a URL for papers for which we did not find a DOI, title, list of authors, year of publication, number of citations in Google Scholar and abstract. Where we had access to the full text, we have also downloaded the PDF. We were able to download 275 PDFs and 296 abstracts. Due to copyright restrictions, the PDFs are not part of the shared dataset[24]. This collection process took a single person several hours a day for about a week.

To obtain readership data, we have used the DOIs, or title and year

[24]As there are Copyright Exceptions for text and data mining in some countries, such as in the UK, we are happy to provide the PDF documents for these purposes to researchers residing in these jurisdictions upon request.

of publication for papers without a DOI, to query the Mendeley API[25]. We were mainly interested in the number of readers of each paper. The dataset contains a snapshot of the Mendeley metadata we were working with. We were able to find 141 out of the 166 seminal papers and 125 out of the 148 literature reviews in Mendeley.

Using the Web of Science (WoS) API[26] we managed to retrieve additional information for the seminal and literature review papers indexed by WoS. We queried the WoS API using publication DOIs, if the document was in the system we obtained a full list of publications citing the paper in question and publications cited by the paper. This list included minimal metadata. In order to get full citation information, we queried the API for each individual (citing and cited) paper.

Finally, we have used the Microsoft Academic (MA) API[27] to obtain additional metadata, as well as citing and cited publications for each paper in the dataset. We have queried the API using publication titles and years.

4.4.4 Dataset analysis

To ensure the collected dataset is suitable for our task, we analyse several statistics describing the dataset including statistics of publication age, distribution across disciplines and citation and readership statistics.

Size

The size of the dataset is presented in Table 4.10. The row *DOIs* shows the number of papers in the dataset for which we were able to find a DOI and the row *DOIs in WoS* how many of these DOIs appear in the

[25] `http://dev.mendeley.com`

[26] `http://wokinfo.com/products_tools/products/related/webservices/`

[27] `http://aka.ms/academicgraph`

Web of Science database. The number of additional references which we collected using the WoS API is shown in the row *Citing & cited references in WoS*, and the number of additional references we collected using the MA API is shown in the row *Citing & cited references in MA*.

The rows *Authors total* and *Unique author names* show the total number of authors of all papers in the dataset and the number of unique author names. To count the unique names, we have compared the surname and all first name initials, in case of a match we consider the names to be the same (e.g. J. Adam Smith and John A. Smith will be counted as one unique name). The *Unique author names* column does not show the number of disambiguated authors, but gives us an indication of how many of the author names repeat in the dataset.

Publication age

Figure 4.10 shows a histogram of years of publication with literature reviews and seminal papers being distinguished by colour. Seminal papers in the dataset are on average about 9 years older than review papers. This shows literature reviews might age faster than seminal papers, which is consistent with our expectations. An explanation for this could be that literature reviews theoretically become outdated as soon as the first new piece of work is published after the publication of the review. Because the seminal papers are on average older this also means these papers had more time to attract citations. This is another reason to expect seminal papers to be distinguishable by citations and readership as features. Descriptive statistics of years of publication both sets are presented in Table 4.11.

Table 4.10: Dataset size.

Responses	171
Seminal papers	166
Review papers	148
Total papers	314
Seminal in Mendeley	141
Review in Mendeley	125
Total in Mendeley	266
Seminal in MA	158
Review in MA	140
Total in MA	298
DOIs	256
Seminal in WoS	48
Review in WoS	58
DOIs (total) in WoS	106
Authors total	1334
Unique author names	1235
Citing & cited references in WoS	19,401
Citing & cited references in MA	153,972

Table 4.11: Descriptive statistics of publication age for both types of papers.

	Seminal	**Review**	**Overall**
Mean	1999	2008	2003
Min	1947	1975	1947
Max	2016	2016	2016
25%	1995	2005	1999
50% (median)	2002	2010	2006
75%	2010	2013	2011

Disciplines

Figure 4.11 shows a histogram of papers per discipline. We have used the information we got about the respondents' academic background to

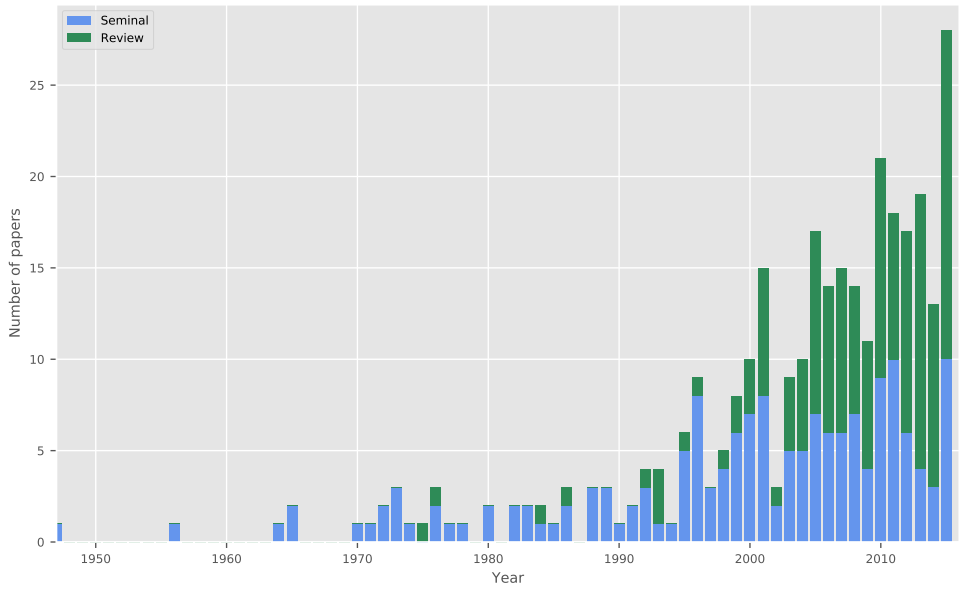

Figure 4.10: Histogram of publication years.

assign papers to disciplines. The respondents have also provided a short description of the research area related to the two references (e.g. "molecular neuroscience", "combinatorics", etc.), however as these descriptions are more detailed and there is little overlap between them we have not used these in our analysis.

The distribution of papers per discipline is to a certain degree consistent with other studies, which have reported Computer Science and Physics to be among the larger disciplines in terms of number of publications, however, Medicine and Biology are typically reported to be the most productive [Althouse et al., 2009, D'Angelo and Abramo, 2015]. The distribution is therefore probably more representative of size of faculties of the Open University than of productivity of scientific disciplines in general, however, we believe this does not influence our study.

When answering the questions about academic background, 22 respondents have selected "Other" instead of one of the listed disciplines, these 22 responses provided us with 40 papers in total. We looked at the

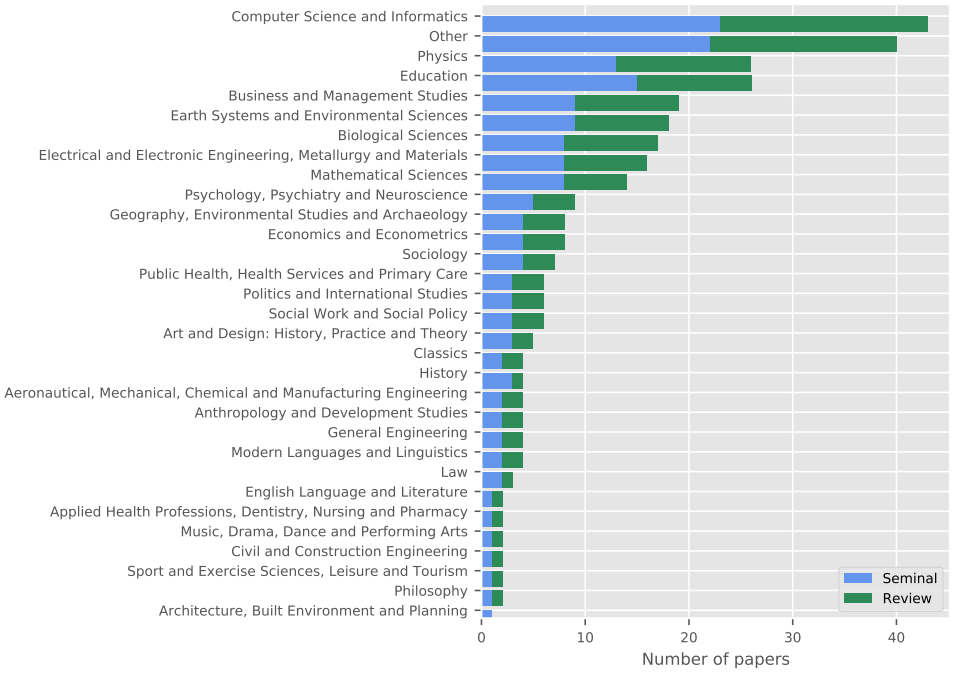

Figure 4.11: Histogram of publication disciplines.

detailed description of these 40 papers, 9 of them are related to astronomy (the descriptions provided were "Binary stars", "Martian meteorites", "cosmochemistry", "Planetary sciences" and "planetology"), 4 could be classified as computer science ("virtual reality" and "Natural Language Understanding, Spoken Language Understanding"), the rest relate to different areas (e.g. "Microbial degradation of plastic" or "MOOC").

Citations and readership

The dataset contains two basic measures related to publication impact – citation counts, which we manually collected from Google Scholar, and the number of readers in Mendeley, which we gathered through the Mendeley API. We also had access to the number of citations in Web of Science and in Microsoft Academic, and while we could not make this data available together with the dataset, we provide an analysis of the

WoS and MA citations, and a comparison with the other two metrics. Table 4.12 shows basic statistics of Google Scholar citation counts and Mendeley readership of each paper in the dataset. We consider the readership of papers which we did not find in Mendeley to be 0 (as papers are added to the Mendeley database by their readers). It is interesting to notice that while seminal papers are on average cited more than review papers, this is not the case for readership, in fact literature reviews attract more readers than seminal papers despite being on average younger. We believe this is an important finding as readership counts are being more and more frequently used as a measure of impact complementary to citations [Piwowar and Priem, 2013, Maflahi and Thelwall, 2016, Priem, 2014]. We believe the fact that literature reviews are more read than seminal papers, while being less cited, suggests that readership can be perceived more as a measure of popularity or utility than a measure of importance.

Table 4.12: Descriptive statistics of Google Scholar citation counts and of Mendeley readership.

	Google Scholar citations			**Mendeley readership**		
	Seminal	**Review**	**Overall**	**Seminal**	**Review**	**Overall**
Mean	2,458	519	1,544	240	368	306
Std	8,885	1,197	6,575	894	1,566	1,264
Min	0	0	0	0	0	0
Max	85,376	12,099	85,376	10,258	15,516	15,597
25%	78	24	41	6	7	7
50%	249	109	194	45	42	46
75%	1,302	596	845	166	145	165

Table 4.13 shows a comparison of citation counts in GS and MA, and Table 4.14 shows a comparison of GS and WoS. The higher citation numbers coming from Google Scholar are not surprising as Google Scholar's

wider coverage of academic outputs is well known [Harzing and Alakangas, 2016, Harzing, 2016]. This wider coverage is also demonstrated by the fact that we were able to find only 298 papers used in our study in MA, and only 106 in WoS.

Table 4.13: Descriptive statistics of citation counts acquired from Google Scholar and Microsoft Academic (MA).

	Google Scholar			**Microsoft Academic**		
	Seminal	**Review**	**Overall**	**Seminal**	**Review**	**Overall**
Count	158	140	298	158	140	298
Mean	2,515	527	1,580	774	227	516
Std	9,085	1,222	6,732	2,048	557	1,561
Min	0	0	0	0	0	0
Max	85,376	12,099	85,276	16,710	5,634	16,710
25%	76	26	42	27	9	13
50%	257	115	195	104	48	84
75%	1,302	596	845	622	203	375

Table 4.14: Descriptive statistics of citation counts acquired from Google Scholar and Web of Science.

	Google Scholar			**Web of Science**		
	Seminal	**Review**	**Overall**	**Seminal**	**Review**	**Overall**
Count	48	58	106	48	58	106
Mean	814	429	607	523	255	379
Std	1,599	566	1,175	926	373	697
Min	2	0	0	1	0	0
Max	8,246	2,446	8,246	4,753	1,709	4,753
25%	102	43	59	46	25	33
50%	211	216	214	144	94	105
75%	929	612	705	677	354	418

This low coverage provided by Web of Science can be seen as a prob-

lem, especially given the fact WoS misses some key seminal papers and overall misses more seminal papers that literature reviews. For example, a recent publication by Krizhevsky et. al. [Krizhevsky et al., 2012], a seminal deep learning paper which has caused a shift in the area of artificial intelligence/computer vision, is missing in WoS, but has (at the time of writing this paper) attracted almost 8000 citations in GS since its publication in 2012. This problem is not limited to WoS either, Scopus for example also does not index the publication, and while Mendeley does, most of the associated meta-data is inaccurate. The most probable reason for these exclusions is that the conference proceedings for this paper are not published through a major publisher but instead by the conference itself and self-hosted on their website. We believe this is an interesting point as it shows important seminal work is not always published by the traditional routes of journals or known publishers. With the recent changes in scholarly communication towards Open Access, Open Science, Arxiv, self hosting, etc. the very definition of "published" no longer has a universal standard and we believe it is reasonable to expect that this will continue with higher frequency as the communities continue to change over time.

In order to compare whether the databases rank papers similarly we have correlated the citation counts (see Tables 4.15 and 4.16). The correlations are in both cases weaker for seminal papers, however this could be caused by the age difference between the two types of papers as the databases might have a lower coverage of older publications. Overall, both Pearson and Spearman correlations are otherwise strong. We believe this shows using citation data from these databases will produce similar results.

Table 4.15: Correlation between Google Scholar and Microsoft Academic citation counts.

	Spearman	**Pearson**
Seminal	0.9010, $p \ll 0.01$	0.4685, $p \ll 0.01$
Review	0.9429, $p \ll 0.01$	0.9830, $p \ll 0.01$
Overall	0.9283, $p \ll 0.01$	0.4941, $p \ll 0.01$

Table 4.16: Correlation between Google Scholar and Web of Science citation counts.

	Spearman	**Pearson**
Seminal	$0.8581, p \ll 0.01$	$0.6775, p \ll 0.01$
Review	$0.9696, p \ll 0.01$	$0.9588, p \ll 0.01$
Overall	$0.9281, p \ll 0.01$	$0.7254, p \ll 0.01$

4.4.5 Summary and discussion

In this section, we have presented a novel dataset of 314 seminal publications and literature reviews for evaluating research metrics, which we made publicly available to the research community. We believe this dataset will be useful in developing and evaluating new metrics. Our goal is to study whether new and existing research metrics distinguish between the two types of papers available in the dataset, and in the next chapter (5) we use the dataset to study citation counts and Mendeley reader counts. While we will show our results are statistically significant ($p < 0.01$) (Chapter 5), a larger dataset would be helpful, especially for studying differences across disciplines. We believe an "ideal" dataset for evaluating research metrics should meet the following requirements:

- **Cross-disciplinary:** A dataset containing publications from different scientific areas is important for two reasons. Firstly, publication patterns are different for each discipline, both in terms

of productivity and types of outcomes (conference papers, journal papers, books, etc.). This is also important to enable detecting research which finds use outside of its domain.

- **Time span:** The dataset should also contain publications spanning a wider time frame. One of the reasons for this is that publication patterns are different not only across disciplines, but they keep changing also in time. Furthermore, some research publications only find use after a certain period of time, but nevertheless represent important research.

- **Publication types:** Different types of research publications (e.g. pure research, applied research, literature review, dataset description, etc.) provide different types of impact. This should be taken into account when developing new research metrics. For example, a publication presenting a system might not receive many citations, because it presents a final product rather than research others can build on. However, such publication might still be widely used and have a large societal or economic impact.

- **Peer review judgements:** Finally, to provide a reference rank for comparing the research metrics to, the dataset should contain fair and unbiased judgements provided by domain experts. These judgements should rate the publications based on an agreed set of rules and standards.

Creating such a dataset would require significant time and resources, both in terms of collecting a representative sample of publications and in terms of providing peer review judgements for these publications. While there was a recent effort to create such a dataset (the dataset created by Microsoft for the 2016 WSDM Cup Challenge, we describe this effort in

Chapter 5), in this case the evaluation set contained only publications from one discipline (computer science) and the peer review judgements were not shared. Providing the peer review judgements could be a common effort and an existing open peer review system could be used for this task. This would require selecting the reference publications, creating a set of rules according to which the papers in the set should be judged and ensuring fairness of the peer review. We believe our study represents the first step in the direction of an ideal evaluation set, as utilising different publication types for metrics evaluation is currently possible. While the creation of such dataset is still time-consuming, it is a less constrained task.

One limitation of our study is that we rely on the respondents' understanding of seminal publications and literature reviews. We have verified the correctness of the responses belonging to the Computer Science and Informatics subset (43 publications), as that is an area most familiar to us. To do this, we have reviewed the publication titles and abstracts. The labelling of this subset matches our understanding of seminal and review publications except in three cases, a paper "From data mining to knowledge discovery in databases" which was labelled as seminal and papers "Process algebra for synchronous communication" and "Unifying heterogeneous and distributed information about marine species through the top level ontology MarineTLO" which were both labelled as a literature review. For these three papers we would flip the labels. We have not however read the full papers and so our disagreement with the respondents could be caused by not knowing the content of the papers and/or not being experts in those areas. As future work we are planning to cross-reference the data to ensure the validity of the entire dataset.

4.5 Conclusions

In this chapter, we have addressed the following question: "How can we evaluate the performance of metrics used in research evaluation for assessing the quality of research publications?" In order to be able to evaluate the performance of an indicator or a metric, two things are typically needed: a sample of research publications to test the metric on, and a ground truth or a validation dataset to compare the metric with to obtain a performance measurement. We have therefore approached this question in two steps.

First, we have reviewed the existing research publication datasets. This review has identified a new dataset, the Microsoft Academic Graph (MAG), which is unique in that it covers most (if not all) disciplines, and contains an openly available citation network. We have provided an analysis of this dataset comparing it with other research publication and citation dataset. We have shown that MAG data correlate well with external datasets, and provide a great resource for research in scholarly communication.

Next, we have reviewed methods which have in the past been used for evaluating research metrics, and analysed the advantages and limitations of each method. The review has shown the lack of a ground truth or a validation dataset which could be used to reliably test new metrics. To help alleviate this issue, we have created a new dataset, which complements the existing approaches, and which consists of 314 seminal publications and literature reviews. The creation of the dataset was based on the idea that these two types of papers provide a very different amount of research contribution and different types of impact. We propose to use this dataset for analyse and evaluate new research metrics. We share

this dataset with the research community[28] and hope it will be useful to others and will perhaps inspire creating a true ground truth evaluation set.

[28] http://trueimpactdataset.semantometrics.org

Chapter 5

Beyond citation counting

> *The only way to discover the limits of the possible is to go beyond them into the impossible.*
> – Arthur C. Clarke

The previous chapter has explored datasets and methods which are typically used for evaluating research metrics, analysed the Microsoft Academic Graph (MAG), which is a new comprehensive research publication and citation dataset, and introduced a new dataset of seminal publications and literature reviews, which can be used for analysing and validating new and existing research metrics. This chapter builds on this previous research work and addresses the following research question:

> RQ3: *What is the relationship between the existing metrics used in research evaluation and the quality of publications?*

More specifically, the aim of this chapter is to analyse the performance of existing (particularly citation-based) metrics for identifying important publications and to investigate whether changes can be made to the existing metrics to improve their performance and make them more robust and reliable. There is an ongoing discussion whether citation counts and

other metrics are appropriate for use in research evaluation. The motivation behind the experiments presented in this chapter is to contribute to this discussion by studying the existing metrics from two different perspectives, but also to create a frame of reference (a baseline) to which new metrics can be compared.

We address our research question in two steps. First, we evaluate the performance of existing research metrics (specifically citation counts and Mendeley reader counts) for distinguishing publications that have caused a change in a research field from those that have not. This experiment has been conducted on a new dataset for bibliometric research, which we call TrueImpactDataset, and which was introduced in the previous chapter (4). We show that citation counts work better than a random baseline (by a margin of 10%) in distinguishing excellent research, while Mendeley reader counts do not work better than the baseline. This gives us a better understanding of the performance of some basic metrics used in the evaluation of research publications and a frame of reference to which we can compare new metrics.

The second part of this chapter is focused on studying whether new methods providing better performance can be designed using the same data the current metrics (especially citations) use. The motivation behind the experiment described in the second part of the chapter is twofold. Firstly, we are interested in studying the performance of existing metrics on a different task (i.e. different than in the first experiment) to broaden and reaffirm our findings from the first experiment. Secondly, given the widespread availability and use of certain metrics (particularly citations), we study whether some improvements and changes to these metrics could be made that would improve their performance without requiring additional data.

The work reported in the second part is based on the research and

results we achieved at an international competition on ranking scholarly publications, 2016 WSDM Cup, organised jointly by Microsoft and Elsevier and associated with the 2016 Web Search and Data Mining Conference (WSDM 2016). WSDM is a major international conference focused on enhancing research in Information Retrieval, Data Mining, and Web Search. The aim of the 2016 WSDM Cup was to assess the importance of scholarly articles using data from MAG [Wade et al., 2016]. This part of the chapter also aims to fulfil Goals 1 and 2:

> Goal 1: *Design new methods for assessing the value of research publications and evaluate these methods in comparison with existing research evaluation metrics.*
>
> Goal 2: *Show that the developed metrics can be deployed in large document collections to improve the analysis of published research.*

For this challenge, we have developed a new method for ranking scholarly publications. The 2016 WSDM Cup has provided an excellent opportunity and framework for experimenting with new research publication ranking methods using the largest publicly available dataset of scholarly publications and citations, the MAG, and the evaluation of these ranking methods in direct competition with methods designed by other teams. As the dataset did not provide abstracts or full text and we could therefore not apply any text-based methods, our focus in this experiment was on analysing existing and new metrics derived from citations.

The content of this chapter is organised as follows. In Section 5.1 we describe our experiment in which we evaluate the performance of citation counts and Mendeley reader counts for distinguishing seminal research publications from literature reviews. In Section 5.2 we present our

experiments focused on evaluating the performance of various citation-based metrics, including citation counts, h-index, and journal impact for ranking publications based on their importance. This evaluation was performed as part of the 2016 WSDM Cup Challenge and used human judgement data as the ground truth. We summarise our findings and conclude the chapter in Section 5.3.

5.1 Do citations and readership identify seminal publications?

This section describes our experiments conducted to evaluate the performance of existing research metrics for identifying important seminal research. In the previous chapter (4), we have reviewed the existing methods for evaluating the performance of research metrics. We have shown a number of different approaches exist, with one of the most common methods being a comparison (typically using correlation analysis) with another metric or metrics. We have demonstrated each of the existing approaches comes with certain advantages and disadvantages, and none help to answer the question completely. To help alleviate this issue, we have created a new dataset, which complements the existing evaluation methods. The dataset, which we call TrueImpactDataset, consists of 314 seminal publications and literature reviews. The idea behind the creation of this dataset was that these two types of papers provide a very different amount of research contribution. As we have shown in Chapter 3, most researchers usually consider the amount of change produced in a field (research contribution, how much did a piece of work move the field forward) to be one of the most important aspects of research publication quality. We use publications which are considered seminal work as examples of research generating a large amount of change in a field and

literature reviews as examples of research typically providing other (such as educational) types of impact. We believe if one of the goals of research evaluation is recognising publications which contributed significantly to their field, seminal papers should perform better under such evaluation than literature reviews, which by definition do not generate a significant change in the field[1].

Therefore, we study how well the existing metrics discriminate between these two types of papers. Our results show that existing metrics help in distinguishing between seminal publications and literature reviews, albeit with room for improvement. We believe this is an important finding demonstrating more attention may need to be paid to publication type in research evaluation, especially as these two types of papers are weighted equally when used in research evaluation metrics such as in JIF [McVeigh and Mann, 2009] and the h-index.

In order to answer our research question, we have designed a simple experiment. We chose citation counts and Mendeley readership as representatives of bibliometrics and altmetrics, as these two measures are both well known and are being used as measures of impact of published research in many settings [Research Excellence Framework, 2012, Wilsdon et al., 2015]. We then classify the papers in the collected dataset into two classes (seminal, review) using two models, a model using the papers' citation counts and a model using their Mendeley readership. We show that the model using citation counts outperforms our baseline by a significant margin, while the model using readership does not perform better than the baseline.

This section is organised as follows. First, in Section 5.1.1 we describe the design of our experiment and the results we obtained using each of

[1]With some exceptions, notably systematic reviews, which are a key practice in evidence-based medicine.

the models. Next, in Section 5.1.2 we provide a discussion of our results. We summarise our findings in Section 5.1.3.

5.1.1 Experiment & Results

In this section, we present the results of the experiment the aim of which was to test whether citation or readership counts work as a discriminating factor for distinguishing seminal papers and literature reviews. These two measures, and especially citation counts, are frequently used as proxies for scientific influence and quality. For example, citation counts are the basis for calculating JIF, where the calculation does not take into account the differences between types of research papers (pure research papers and literature reviews are both used as input with equal weight) [Thomson Reuters, 2012]. As we have shown in Chapter 3, amount of research contribution is often indicated as an important dimension of research quality [Research Excellence Framework, 2012, Tertiary Education Commission, 2013b, Australian Research Council, 2015a]. Thus, we study how well do these two types of papers distinguish between publications generating very different amounts of research contribution.

In order to test our hypothesis we use these two metrics to classify the papers into the two classes (seminal, review). As a baseline we use a model which classifies all papers as seminal, as that is the majority class. This baseline model achieves the accuracy of 52.87%. We calculate accuracy as the proportion of correctly classified publications, or more formally:

$$acc = \frac{TP + TN}{N} \tag{5.1}$$

where the category *seminal* is our positive class, TP (true positives) is the number of items correctly labelled as belonging to the positive

class, TN (true negatives) is the number of items correctly labelled as not belonging to the positive class, and N is the number of all items (publications).

Before running the experiments we first perform a statistical test to see whether the citation/readership distributions of seminal and review papers differ. We perform a one-tailed independent t-test with the null hypothesis stating that the means of the two groups are equal. The results we get are $p = 0.0063$ for citations and $p = 0.1666$ for reader counts. In case of citations, for a significance threshold of 1% we reject the null hypothesis. Because we know the mean number of citations of the seminal papers is higher (Table 4.12), we conclude seminal papers are cited significantly more than literature reviews. In case of readership, we accept the null hypothesis that the distributions of reader counts of seminal and review papers are the same (that is the number of readers does not distinguish between the two groups). To better understand how well each metric works in distinguishing between the two groups, we use citations and readership as features in a classification experiment.

The classification experiment relies on two approaches. First, we use a leave-one-out cross-validation setup, that is we repeatedly train on all but one publication and then test the performance of the model on the publication we left out of the training. We do this for all publications in the set. However, in some cases, due to the size of the dataset, leaving out even one publication can affect the performance of the model. For this reason we also find the performance of the ideal model, that is we train the model on all available data. This gives us an upper bound of performance. We run three separate experiments. First, we train and test our models on all available data. This gives us an idea of how well do both metrics perform across disciplines and regardless of time. We call this the aggregate model. Next, we split the data by discipline and

create separate models for each discipline. Finally, we split the data by publication years and create separate models for each year. It would be interesting to also split the data by both discipline and year, however, we were not able to do this due to the size of the dataset, as the resulting groups would be too small for analysis.

Aggregate model

The model we use to classify papers based on their citation and reader counts works in the following way: if the total number of citations (or the number of readers) for a given paper is equal to or greater than a selected threshold we classify the paper as seminal, otherwise as a literature review. To do this, we use the threshold which achieves the best accuracy (which is calculated as the number of correctly classified examples divided by the number of all examples) on the training data. We find this threshold by calculating the accuracy for all thresholds in the interval $[0, max(citation\_count)]$ for the model using citation counts and $[0, max(reader\_count)]$ for the model using reader counts. If there is more than one such threshold, we use the average value of all best thresholds. For the ideal model we chose any of the best thresholds, as all will have the same performance.

Table 5.1 shows the confusion matrix for the leave-one-out cross-validation scenario using **citation counts** as a feature. This setup achieves an overall accuracy of 63.06%, which represents about 10% improvement over the baseline. All but two of the models trained in the cross-validation setup chose 51 citations as an optimal threshold (the two other thresholds were 52.4 and 52.5). The ideal model (trained on all available data) achieves the accuracy of 63.38%.

Furthermore, Table 5.2 shows the confusion matrix obtained by using **reader counts** as a feature. This model achieves an overall accuracy of

42.68%, which is about 10% worse than the baseline. Most of the models (277) trained in the cross-validation setup chose 0 readers as the optimal threshold. The remaining models (37) chose 2.5 readers as a threshold. In this case, the performance of the ideal model is 52.87%, which is equal to the baseline.

Table 5.1: Confusion matrix for predicting the class of the paper using Google Scholar citation counts.

		Predicted		
		Review	**Seminal**	**Total**
Actual	**Review**	19.43% (61)	27.71% (87)	148
	Seminal	9.24% (29)	43.63% (137)	166
	Total	90	224	314

Table 5.2: Confusion matrix for predicting the class of the paper using Mendeley reader counts.

		Predicted		
		Review	**Seminal**	**Total**
Actual	**Review**	0.00% (0)	47.13% (148)	148
	Seminal	10.19% (32)	42.68% (134)	166
	Total	32	282	314

Discipline based model

This model uses discipline information to first split the papers into groups. For all separate groups we then perform the same statistical test and classification experiment using both citation and reader counts. In this case, we remove all papers labelled as "Other". Furthermore, we remove all subject areas which contain less than two of each type of papers, to be able to train and test the models on representatives of both seminal and

review papers. The p-value is greater than 1% for all remaining disciplines and for both citation and reader counts, which means in all cases we accept the null hypothesis of equal averages. All p-values are shown in Appendix C, Table C.1.

The overall cross-validation accuracy is 45.28% for citations and 42.13% for reader counts, which is worse than the baseline (52.87%) in both cases. We believe this is due to the fact the baseline is not dependent on the size of the data, while in the leave-one-out cross-validation, removing even one paper can change the performance of the model. Furthermore, the baseline method "knows" which class is the majority class, while our model does not use this information. Both of these factors make it harder to outperform the baseline. The results for separate disciplines are reported in Appendix C Tables C.2 and C.3. To calculate the overall accuracy, rather than counting average accuracy across all disciplines, we sum all confusion matrices and calculate the accuracy from the sum (Tables 5.3 and 5.4, this method is sometimes referred to as *micro-averaging*). The accuracy of the optimal model goes up in both cases, to 68.11% in the case of citations and to 62.60% in the case of readership. This shows that separating papers by discipline has the potential of improving the results.

Table 5.3: Overall classification results obtained from running the classification for each discipline separately, using citations as a feature.

		Predicted		
		Review	**Seminal**	**Total**
Actual	**Review**	24.41% (62)	23.62% (60)	122
	Seminal	31.10% (79)	20.87% (53)	132
	Total	141	113	254

Table 5.4: Overall classification results obtained from running the classification for each discipline separately, using reader counts as a feature.

		Predicted		
		Review	**Seminal**	**Total**
Actual	**Review**	17.32% (44)	30.71% (78)	122
	Seminal	27.17% (69)	24.80% (63)	132
	Total	113	141	254

Year based model

We perform a similar experiment as in case of disciplines also for publication years. We split the publications in the dataset into groups by the the year in which they were published and again leave out those groups which do not contain at least two papers of each type. The p-value is greater than 1% for all publication years (Table C.4 in Appendix C). The overall cross-validation accuracy is 55.23% (Table 5.5) for citation counts and 51.05% (Table 5.6) for reader counts, which in the case of citation counts is an improvement both over the baseline (52.87%) and over the previous model trained per discipline. The accuracy of the optimal model is 68.62% in the case of citations and 65.27% in the case of reader counts. The full results are reported in Appendix C, Tables C.5 and C.6.

Table 5.5: Overall classification results obtained from running the classification for each year separately, using citations as a feature.

		Predicted		
		Review	**Seminal**	**Total**
Actual	**Review**	39.75% (95)	17.15% (41)	136
	Seminal	27.62% (66)	15.48% (37)	103
	Total	161	78	239

Table 5.6: Overall classification results obtained from running the classification for each year separately, using reader counts as a feature.

		Predicted		
		Review	**Seminal**	**Total**
Actual	**Review**	37.66% (90)	19.25% (46)	136
	Seminal	29.71% (71)	13.39% (32)	103
	Total	161	78	239

5.1.2 Discussion of results

Table 5.7 shows a summary of classification results of all three models. The year based model performs better than the discipline based model, however this might be due to the distribution of survey and seminal publications in our dataset – as we have shown in Chapter 4, Table 4.11, seminal papers in our dataset are on average older than literature reviews, which makes the year based classification easier. In reality papers published in a given year will be distributed more evenly. The performance of the discipline based model should be more stable, as the distribution of seminal and survey papers across disciplines in our dataset is more even. We have not performed a classification across both disciplines and years as due to their wide distribution we were not able to find enough examples belonging to the same discipline and year. The aggregate model outperforms the two other models, however, we believe this might be due to the size of the dataset. The accuracy of the ideal models suggests splitting the publications both by discipline and by year has the potential of improving the results.

5.1.3 Summary

There has been much discussion on whether citation counts are appropriate for use in evaluation of research outputs [Wilsdon et al., 2015].

Table 5.7: Summary of all results. Column *Accuracy* shows the accuracy obtained in the leave-one-out cross-validation scenario, while column *Ideal acc.* shows a theoretical upper bound of performance (an accuracy of a model trained on all available data).

Model	**Data**	**Accuracy**	**Ideal acc.**
Baseline	Citations	-	52.87%
	Readership	-	52.87%
Aggregate	Citations	63.06%	63.38%
	Readership	42.68%	52.87%
Discipline based	Citations	45.28%	68.11%
	Readership	42.13%	62.60%
Year based	Citations	55.23%	68.62%
	Readership	51.05%	65.27%

We have used a new approach to study this question. Specifically, we studied how well citation counts and Mendeley reader counts distinguish important seminal publications that have changed a research field from publications that have not. We have performed a set of experiments using citation and reader counts to classify papers into seminal and literature review categories and showed that citation counts help in distinguishing important seminal research from literature reviews with a degree of accuracy (63%, i.e. 10% over a random baseline), while Mendeley reader counts don't work better than a random baseline on this task and our dataset (highest accuracy 51.05%, while our baseline model achieved 52.87%). This contributes to answering our Research Question by demonstrating on a real dataset of research publications that citation counts to a certain degree work as a research metric for assessing research contribution, albeit with a room for improvement. Our results show that caution should be exercised when using citation counts for certain tasks (even if discipline and age is taken into account). We believe that while cita-

tions seem to work to some degree, additional methods, such automated methods for classifying important citations [Teufel et al., 2006, Valenzuela et al., 2015, Pride and Knoth, 2017], may be needed to further improve the performance of these metrics.

5.2 Simple yet effective methods for large-scale scholarly publication ranking: KMi and Mendeley (team BletchleyPark) at WSDM Cup 2016

In the previous section we have shown citation counts work with a degree of accuracy (63%, i.e. 10% over a random baseline) as a metric for assessing the amount of research contribution of a publication. In this section we present the results of a further evaluation of the performance of citation counts and present a new simple publication ranking method with significantly better performance in our task than simple citation counts. This evaluation has been conducted through participation in the 2016 WSDM Cup challenge, in which the submitted publication ranking methods were evaluated against human judgement data [Wade et al., 2016]. The participation in the challenge has therefore enabled us to evaluate the performance of citation counts (including normalised citation counts, the h-index, and other related metrics) against data, which is otherwise difficult to obtain. As the dataset used in the challenge did not provide publication abstracts or full text we were unable to experiment with any text-based approaches. In this experiment, we have therefore focused our attention on evaluating and extending the existing citation-based metrics.

The goal of the challenge was to assess the importance of research pub-

lications using data from the Microsoft Academic Graph (MAG, Chapter 4) and to provide a static rank for publications in the dataset. The submissions to the challenge were scored based on agreement with human judgement data (which were provided by experts in the field) on a subset of Computer Science publications [Wade et al., 2016]. The judgement data were randomly split into an evaluation and a test set, and the challenge was done in two phases.

During the first phase, the submissions of the participating teams were scored automatically against the evaluation set, and the score was displayed on a public leaderboard[2]. During this phase, the participating teams were allowed to upload any number of submissions and to test different ranking methods against the evaluation set. After the end of the first phase, the teams were no longer able to upload new submissions, and their most recent submission was used to evaluate each team against the test set [Microsoft Research, 2015]. This two step evaluation was intended to prevent teams from overfitting their methods to the evaluation set. After this first round of the challenge, the top eight teams were invited to re-run their methods on an updated version of the dataset and submit new rank values. During the second phase of the challenge, the eight winning teams were evaluated through the Bing[3] search engine, in which the submissions were used to rank publications for academic search queries.

Our approach to the challenge was based on the assumption that the importance of a publication can be determined by a mixture of factors evidencing its impact (factors directly related to the publication) and the importance of entities which participated in the publication's creation

[2]The leaderboard, which currently displays the performance of the eight winning teams on the test set, was available at `https://wsdmcupchallenge.azurewebsites.net/Home/Leaderboard`

[3]`http://www.bing.com/`

(factors related to the authors, venue, etc.). Our method has achieved encouraging results (it ranked fist on the evaluation set and fifth on the test set, compared to methods submitted by over 30 participating teams), and we describe in detail how the performance can be further improved.

This section is organised as follows. We start by presenting our ranking method (Section 5.2.1). In Section 5.2.2 we discuss the performance and potential improvements to our method. In Section 5.2.3 we provide a discussion of our results and of the evaluation method used in the challenge. In Section 5.2.4 we review the ranking methods submitted by the other finalists. Finally, in Section 5.2.5 we summarise our findings.

5.2.1 Publication Ranking Methods

The task and the data

The 2016 WSDM Cup Challenge can be described as follows: given a heterogeneous graph, which models real-life academic communication, find a static rank value for each publication entity in the graph representing the papers' importance in the graph. Our approach to solving this task is in detail described in the remainder of this section.

In the 2016 WSDM Cup Challenge the performance of different methods was assessed on the MAG dataset (Chapter 4), which consists of six types of entities: scholarly publications, authors, institutions, fields of study, venues (journals and conferences, e.g. WSDM) and events (specific conference instances, e.g. WSDM 2016). The dataset also contains citation relationships between the publication entities. A detailed description of the entities and their relationships is provided in [Sinha et al., 2015]. We have also presented a detailed analysis of the dataset with focus on the utility of the dataset for research evaluation in Chapter 4.

Our approach

Our approach was based on the hypothesis that the importance of a publication can be determined by a mixture of factors evidencing its impact and the importance of entities which participated in the publication's creation. We believe method transparency is an important characteristic, for this reason we were trying to come up with a simple, understandable and transparent method which could potentially improve the current situation in research evaluation. The approach used in our submission was based on the following method. We have separately scored each of the types of entities in the graph (we have produced a separate score for authors, institutions, journals, etc.). We have then used the separate scores to provide a publication score (e.g. we have scored publications based on the scores of their authors, or based on the venue at which they were published). In this way we have produced several different scores for the publication entities. The final score, which determines the publication's rank among its peers, was then calculated using linear combination of these scores. We have experimented with different combinations of different methods presented in this section, as well as different weights. The standard approach for determining weights for the separate scores would be to use machine-learning approach, however because no ground truth data were available for training and verifying the methods, we deduced the weights experimentally. Equation 5.2 shows the final weights. This equation was used to produce our final submission for the second round of the challenge.

$$\begin{aligned} score(p) =& 2.5 \cdot s_{pub} + 0.1 \cdot s_{age} + 1.0 \cdot s_{pr} + \\ & 1.0 \cdot s_{auth} + 0.1 \cdot s_{venue} + 0.01 \cdot s_{inst} \end{aligned} \tag{5.2}$$

The differences between our first and second round submissions, each

of the separate ranks as well as which alternatives did we experiment with are described in the remainder of this section.

Publication-based scoring functions

To score the publication entities directly, without considering the score or importance of their authors or venues, we have utilised the citation relationships provided in the graph. The simplest option is to score the publications solely by the number of citations they receive. We have experimented with several options of normalising and weighting the citations, namely:

Applying a time decay to citations. We have used an exponential decay function $f(t) = e^{-\alpha(t_c - t)}$, where t_c is the current year, t is the year in which the paper from which the citation originates was published and α is a constant influencing the decay rate. This means that each citation contributes to the total fully only in the year in which it originates, and the value of the citation diminishes with age. The rationale behind this is to distinguish between publications which received attention only years after publication and those which are still presently used [Del Corso and Romani, 2009]. We have experimented with several different values of α, ranging from 0.05 (slower decay) to 0.15 (faster decay).

Applying a decay function to total citation counts. The idea behind applying a decay function to the citation total is that the importance of publications does not necessarily increase linearly with the increasing number of received citations. For example, it has been suggested that the concept called the *Matthew effect*, where highly cited papers (as well as researchers, etc.) receive a cumulative advantage, could be at work in science [Merton, 1968, Price, 1976]. We have experimented in using logarithmic and linear decay, however we have achieved the best results when simply setting a maximum threshold for the total citation count

above which the received citations are no longer considered.

Using normalised citation counts. Normalising total citations to citations received per year since the publication of the paper, per author of the paper, and per year and author. It has been suggested that the number of authors on the paper could cause a multiplication effect of specific audiences for each involved author [Bornmann and Leydesdorff, 2015]. The use of citations per year is a simplification of the time decay function.

We have found the total number of citations per author of the publication with maximum threshold for the citation total to perform the best. We write this part of the equation as follows:

$$s_{pub}(p) = \begin{cases} c(p)/|A_p|, & \text{for } c(p) \leq t \\ t/|A_p|, & \text{for } c(p) > t \end{cases} \tag{5.3}$$

where $c(p)$ is the total number of citations received by p, A_p is the set of authors of p and t is the threshold. We have experimentally set the threshold to $t = 5000$. This version of the equation is a slightly updated version for the second round of the challenge. In the first round, the second part of the equation was defined as $0/|A_p|$, for $c(p) > t$.

Furthermore, to account for publication age, we use a score based on the age. This score is a simple linear function of publication age and can be written as

$$s_{age}(p) = y_p \tag{5.4}$$

where y_p is the year of publication of p. Based on this score, papers published in the current year have the highest importance and as time elapses their importance linearly decreases.

In the second phase of the WSDM Cup Challenge we have also computed the PageRank [Brin and Page, 1998] value for each of the public-

ation entities in the graph. To allow for efficient PageRank calculation, we chose an approach similar to [Bini et al., 2008] and introduced a new "dummy" paper in the network, which is cited and cites all publications in the citation network except for itself. This paper collects and redistributes weight equally to all publications in the network. This part of the equation can be written as

$$s_{pr}(p) = PR(p) \tag{5.5}$$

We have found the PageRank score to perform similarly to total citation counts and we added the PageRank value as an additional feature. Table 5.8 shows scores we obtained with each of the tested alternative publication ranking methods separately. We have also experimented with different variants and combinations of the methods listed in the table, however, the listed methods obtained the highest scores. According to the organisers, the scores (which we obtained from the public leaderboard after submitting our results) were calculated using Pairwise Correctness [Wade et al., 2016] (more information about the scoring function was not provided).

Author-based score

Commonly used methods for evaluating author performance include the total number of citations received by an author, average number of citations per author's publication and indices such as the h-index [Hirsch, 2005]. We have experimented with these three methods. We calculated the given value for each of the authors of a publication and then tested ranking the publication entities using the maximum, total and mean of the values of the publication's authors (e.g. using maximum, total and mean of the authors' h-index values). We found the mean value of citations per author's publication to perform the best. The author-based

Table 5.8: Scores obtained during the evaluation phase using different publication ranking methods based on publication information. For comparison, we have also included a score obtained by ranking publications using random numbers.

Method	**Score**
Total number of received citations	0.687
Total citations with exponential time decay, $\alpha = 0.05$	0.701
Total citations with exponential time decay, $\alpha = 0.10$	0.705
Total citations with exponential time decay, $\alpha = 0.15$	0.703
Number of citations normalised by publication age	0.695
Total number of citations divided by number of authors	0.699
Our final $s_{pub}(p)$ ranking function	0.711
Random rank	0.024

rank we used can then be expressed as

$$s_{auth}(p) = \frac{\sum_{a \in A_p} \frac{\sum_{x \in P_a} c(x)}{|P_a|}}{|A_p|} \tag{5.6}$$

where P_a is a set of publications authored by a. Table 5.9 shows scores we obtained by ranking the publications in the dataset using each of the tested author-based ranking methods.

Venue-based score

The metric which is considered the standard in journal evaluation is the Journal Impact Factor (JIF) [Garfield, 1972]. The JIF calculation concerns the computation of a mean number of citations received per item published in the journal during a specified time frame, typically during two years prior to the current year. Alternative journal evaluation metrics include the Scimago Journal Rank [González-Pereira et al., 2010] and the Eigenfactor [Bergstrom, 2007] which both revolve around the idea

Table 5.9: Scores obtained during the evaluation phase using different ranking methods based on available author information.

Method	**Score**
Most cited author of the authors of p	0.558
Sum of citations of all authors of p	0.576
Sum of citations of all authors of p divided by number of authors	0.588
Maximum value of h-indices of all authors of p	0.504
Sum of h-index values of all authors of p	0.550
Mean h-index value across all authors of p	0.570
Our final $s_{auth}(p)$ ranking function	0.667

that citations from high-impact journals provide a larger contribution to the importance of a journal than citations from poorly ranked journals.

In evaluating conferences no established metric similar to JIF or other journal evaluation metrics exists. However, a similar approach as in case of journals can be used also for evaluating conferences. We have experimented with few simple scoring functions, such as with total number of citations received by a venue and mean number of citations per paper published at the venue, and with applying these scores to the papers published at the venue (this is an approach similar to the JIF, however we have used all papers published during the existence of the journal or conference). Our final venue-based score can be calculated as

$$s_{venue}(p) = \sum_{x \in P_v, x \neq p} c(x), \tag{5.7}$$

where P_v is a set of papers published at a venue v. Table 5.10 shows scores we obtained by ranking the publications in the dataset using different methods.

Table 5.10: Scores obtained during the evaluation phase using different publication ranking methods based on venue information.

Method	**Score**
Total venue citations	0.159
Mean venue citations	0.159
Our final $s_{venue}(p)$ ranking function	0.341

Institution-based score

Various approaches exist to evaluating institutions. The Nature publishing group ranks institutions based on the number of articles published in their journal Nature[4]. Scimago Institution Rankings[5] provide a list of indicators, including the total number of documents published in scholarly journals, proportion of highly cited publications and rate of collaboration with foreign institutions. In our approach we have however used a simple method similar to the author and venue score. Our final institution-based score can be expressed as

$$s_{inst}(p) = \frac{\sum_{i \in I_p} \sum_{x \in P_i, x \neq p} c(x)}{|I_p|} \tag{5.8}$$

where I_p is a set of (unique) institutions of the authors of the publication and P_i is a set of publications published by authors affiliated with institution i. Table 5.11 shows scores we obtained by ranking the publications in the dataset using different institution-based ranking methods.

[4] `http://www.natureasia.com/en/publishing-index/global/`

[5] `www.scimagoir.com/`

Table 5.11: Scores obtained during the evaluation phase using different publication ranking methods based on institution information.

Method	**Score**
Sum of all citations received by all affiliated institutions of p	0.418
Sum of mean institution citations	0.414
Sum of mean institution citations, divided by number of institutions	0.412
Our final $s_{inst}(p)$ ranking function	0.512

5.2.2 Experiments

Final performance

We have experimented with different combinations of the methods presented in the previous section as well as different weights. During the training phase of the challenge we have submitted over 270 runs. The final score we have obtained at the end of the first phase using Equation 5.2 as our ranking function (after finding the optimal weights and a combination of methods) was 0.769 on the evaluation set, and 0.659 on the test set. Specifically, the ranking function used at the end of the first phase consisted of five separate ranking functions which were combined into a final rank using a weighted sum: a ranking function s_{pub} based on publication information (citations), a ranking function s_{age} based on publication age, a ranking function s_{auth} based on author information, a ranking function s_{venue} based on venue information, and a ranking function s_{inst} based on institution (affiliation) information. The specific ranking functions used were presented in the previous section. The weights used to produce the final rank were as follows:

$$\begin{aligned} score(p) =& 2.5 \cdot s_{pub} + 0.1 \cdot s_{age} + 1.0 \cdot s_{auth} + \\ & 0.1 \cdot s_{venue} + 0.01 \cdot s_{inst} \end{aligned} \tag{5.9}$$

In the second phase of the challenge we have additionally computed the PageRank value for each of the publication entities in the graph and added the value to Equation 5.9 with the weight of 1.0, i.e. $score_{r2}(p) = score(p) + 1.0 \cdot s_{pr}$.

Evaluation

According to the organisers the submitted results were evaluated based on the percentage agreements with human evaluation data [Microsoft Research, 2015], using Pairwise Correctness as the evaluation metric [Wade et al., 2016]. The evaluation data were prepared by Computer Science experts who conducted pairwise ranking of a subset of the MAG dataset. The evaluation data have then been split into validation and test set. While the challenge was running, the participants could evaluate their results against the test data through an online evaluation tool, which provided a score for each of the submitted runs. At the end of the first round of the challenge, the last submitted run of each team was scored against the validation set.

Performance comparison with other teams

The performance of all participating teams was provided both during and after the first round of the challenge through a public leaderboard. According to the leaderboard ranks, our method has achieved the highest score on the test data, and has been ranked as fifth best when scored against the validation data [Microsoft Research, 2015].

Potential improvements

There is a number of ways in which our method could be improved. We believe the main possibilities include the following options.

Better utilisation of the citation network. Due to resource limitations, we were only able to compute PageRank of the publication entities later in the challenge. We see a potential improvement in computing additional network measures, such as different centrality indices, for all entities in the graph.

Inclusion of additional data sources. At the beginning of the challenge we explored the possibility of obtaining additional data. In particular we were interested in utilising altmetric [Galligan and Dyas-Correia, 2013] and webometric [Almind and Ingwersen, 1997] data sources and acquiring publication full-texts or abstracts for use in semantometric measures (Chapter 6). For altmetric and webometric data we have investigated the feasibility of obtaining data from Altmetric.com, Mendeley, ResearchGate, ImpactStory, and ArXiv. For the publication full-texts we have investigated Elsevier, Springer, CrossRef, and Mendeley APIs. Unfortunately most of the investigated services either did not provide an interface for downloading all of their data, or their coverage was too low, which is why we eventually dropped this idea. However, particularly if access to the publication full-texts was possible, this option could provide valuable additional information, for example by extending simple citation counts to research contribution (Chapter 6). A more detailed discussion of the alternative methods is provided in Section 5.2.3.

Possibility to analyse the evaluation data and metric. It is not clear if and up to what extent do the expert judgements correspond with the importance of the publications. Publishing the evaluation dataset and details of the evaluation metric would help in understanding whether the methods submitted to the challenge could help in improving user experience and research evaluation. However, the challenge organisers chose to not share the evaluation data.

Revise the maximum citation threshold used the s_{pub} score. Because

the evaluation data were not shared, we were not able to determine why this threshold led to the improvement of our results.

5.2.3 Discussion

What have we learned

In scoring each of the graph entities we have experimented with different options, from simple citation counts to applying decay functions, calculating PageRank and h-index. It is interesting that in each case, a method based on total or normalised citation counts produced better results than using these widely used measures. Regardless of whether better scoring functions can be found, we believe that in order to develop a more optimal ranking method, it is crucial to better understand the evaluation data and method (what is required from the ranking system). Although a simple approach based on citation counts produced the best results, this does not mean such method will work equally well in real-life settings. For example, it is not clear how much are the human judgement data biased towards citation counts. This issue could manifest in case the judges had access to such information when rating the publications. Furthermore, although citation counting provides a simple and easily understandable ranking method, as we have shown in Chapter 1, it does not account for many characteristics of citations, including the differences in their meaning [Nicolaisen, 2007], popularity of certain topics and types of research papers [Seglen, 1997], the skewness of the citation distribution [Seglen, 1992] and the time delay for citations to show up [Priem et al., 2010].

Evaluation

The goal of the 2016 WSDM Cup challenge was to assess the importance of scholarly articles while exploring alternatives to citations. The format of the results was in the 2016 WSDM Cup defined as a ranked list of the MAG publication entities. In order to evaluate these results, the evaluation setup consisted of the evaluation data – reference ranks prepared by human judges – and an evaluation metric. While preparing our submission, we have identified few problems of the evaluation setup. One of these problems, which we discussed in the previous paragraph, is the subjectivity of the evaluation dataset. While the description of the task encouraged exploration of approaches alternative to citations, it was not clear whether the evaluation setup was capable of potentially rewarding properties of such approaches. Our citation-based method has achieved a high score. Furthermore, due to the fact that the details of the evaluation data were not shared, it became more complicated to avoid overfitting our model. The availability of a good evaluation framework is crucial for enabling the development of new ranking methods and for comparing different approaches. We believe a good evaluation framework should favour properties of the desired ranking system, and the method of creation of this dataset should be transparent to facilitate understanding any biases present in the dataset and to help preventing overfitting.

Alternative ranking methods

In section 5.2.2, we list the external datasources which we investigated. Our motivation for exploring these external datasources was the hope of utilising new altmetric and webometric research evaluation methods. The advantage of these approaches lies for example in the early availability of the required data, when compared to the delay with which citations

show up. These metrics also provide a broader view of publications' impact. However, our main interest lies in the utilisation of publication full text for research evaluation (Chapter 6). The biggest problem of utilising full-text is the difficulty of obtaining the full texts due to various copyright restrictions and paywalls. The MAG dataset could be a very valuable resource for further research if it could be combined with publication full texts, and altmetric and webometric datasets. An interesting future direction could be to enrich the MAG with these data and organise another run of the challenge with the possibility to use these additional data.

5.2.4 Review of solution submitted by other teams

Out of the 32 participating teams, the top eight teams were invited to participate in the second phase of the challenge and present their solution at the 2016 WSDM Cup Workshop. As one of the eight top teams chose not to participate in the workshop, here we review the solutions submitted by the remaining teams. Table 5.12 shows final scores achieved by the seven winning teams who presented their solutions at the workshop (the team which did not participate placed seventh).

Table 5.12: Final scores of the seven top teams obtained on the test set.

Rank	**Team**	**Score**
1	[Feng et al., 2016]	0.6838
2	[Wesley-Smith et al., 2016]	0.6760
3	[Ribas et al., 2016]	0.6713
4	[Hsu et al., 2016]	0.6636
5	Our solution	0.6589
6	[Luo et al., 2016]	0.6558
8	[Chang et al., 2016]	0.6417

Five of the presented solutions ([Feng et al., 2016, Wesley-Smith et al., 2016, Hsu et al., 2016, Luo et al., 2016, Chang et al., 2016]) were based on a variation of the PageRank algorithm [Brin and Page, 1998], while the remaining solution ([Ribas et al., 2016]) used a simplified version of the Relative Citation Ratio metric [Hutchins et al., 2016]. Feng et al. [2016] utilised paper, author, and venue entities available in the graph. They first assigned a score to all papers, which was based on a linear combination of number of citations (how many papers cite a given paper) and number of references (how many papers a given paper cites). The then iteratively performed score propagation and refinement steps. In the score propagation step, paper scores were propagated to author, venue, and cited publication entities. In the score refinement step, venue scores were propagated to authors, and paper scores were recalculated using the author, venue, and citation scores. The new paper scores were then carried over to the next iteration, while the author and venue scores were reset.

Luo et al. [2016], Chang et al. [2016] and Hsu et al. [2016] introduced time into their models. The underlying idea for incorporating time into the ranking models is that the importance of a publication gradually decreases as it becomes older. Luo et al. [2016] utilised information about citation peak time (period during which an article receives the most attention) and decreased the weight of citations to an article after the peak. They then used these weighted citation edges to produce and rank several different graphs (citation, venue, author, and affiliation graphs), which are afterwards used to produce a final ranking for publications using a weighted combination of the separate ranks. Hsu et al. [2016] and Chang et al. [2016] on the other hand used publication age to produce the initial paper weight by calculating the average number of citations per year. Hsu et al. [2016] used these paper weights to calculate venue,

author, and affiliation weights, which were then summed to produce new paper weights. Chang et al. [2016] used a similar propagation method as Luo et al. [2016], but utilised a variant of the HITS algorithm instead of PageRank to calculate hub scores for authors, conferences, and papers. The hub scores are then used to calculate authority scores for papers.

Wesley-Smith et al. [2016] utilised a version of the Eigenfactor metric [Bergstrom, 2007] called Article-Level Eigenfactor (ALEF) for ranking scholarly articles and an extended version of ALEF for ranking authors. Both metrics work similarly as PageRank by simulating a random walk on the citation network. The final score was computed using a weighted sum of the paper and author scores.

An interesting approach was chosen by Ribas et al. [2016] who utilised a simplified version of the Relative Citation Ratio (RCR) metric [Hutchins et al., 2016]. The difference between the original RCR metric and the simplified version, referred to as S-RCR by Ribas et al. [2016], is in the normalisation step. While the original metric utilises linear regression of a co-citation neighbourhood of a paper to perform normalisation, the simplified version uses an average value of a paper's neighbours to perform normalisation. The authors used additive smoothing to overcome situations when a publication has no co-citation neighbourhood. This single feature has performed very well and has scored third on the test set (Table 5.12).

5.2.5 Summary

In this section we presented our method for assessing the importance of scholarly publications, which we submitted to the 2016 WSDM Cup Challenge. Our method was ranked among the top performers in the challenge. We have presented several potential improvements to the method and the knowledge acquired when carrying out experiments. Our find-

ings highlight the difficulty of progressing beyond citation counts. While MAG is an extremely useful dataset for testing evaluation metrics, it would be extremely valuable if this dataset was merged with other sources evidencing impact, as this would enable developing and testing fundamentally new metrics. Additionally, there is a need for a large, open, and unbiased dataset of human judgements to move us closer to this goal.

5.3 Conclusion

In this chapter, we have addressed the following question: "What is the relationship between the existing metrics used in research evaluation and the quality of a publication?" In order to answer this question, we have evaluated the existing metrics on two datasets and using two different methods. First, we have studied the performance of citation and Mendeley reader counts in distinguishing publications that have changed a research field from those that have not. This evaluation has shown that citation counts help in distinguishing these two types of papers with a degree of accuracy (63%, i.e. 10% over a random baseline), while Mendeley reader counts do not distinguish between these two types of papers at all.

Next, we have evaluated the performance of citation counts and several related metrics, including the h-index and the journal impact factor, for ranking scholarly publications according to their importance. In this evaluation, the submitted methods were compared to ranks produced by domain experts. In our experiments we have made several interesting observations. For example, ranking publications solely based on citation counts received by their authors has worked fairly well and performed only a little worse than ranking publications using number of times they were cited. On the other hand, ranking publications based

on the venue in which they appeared did not perform very well, but it improved the performance of our method when used in combination with publication and author information. Furthermore, ranking publications using h-index values of their authors performed slightly worse than using a simple mean number of citations per author. This suggests overall author performance may be more important than the performance of their top cited publications.

We have demonstrated that by combining the information from different types of entities (publications, authors, venues, and affiliations) even without utilising additional data such as text, we can achieve significantly better performance than by utilising information from a single type of entity at a time. We believe this is an important finding, as it demonstrates simple improvements can be made to the existing research metrics to improve their performance. One limitation of this experiment is that the evaluation was performed on human judgement data. As the description of the evaluation data provided by the challenge organisers was not very detailed, it is possible that the judgements were potentially biased or inaccurate. However, at the time of the challenge and our evaluation, this evaluation represented a state-of-the-art method, and no better alternative was available.

Chapter 6

Semantometrics: Towards content-based research evaluation

Absence of evidence is not evidence of absence.

– Carl Sagan

In the previous chapter we have analysed the performance of existing research evaluation metrics in two separate tasks – assessing the importance of research publications to produce rankings similar to those produced by human experts and distinguishing important seminal publications from literature reviews. We have shown that while simple citation counts as well as normalised values work to a certain degree, we can achieve a significant performance improvement by combining ranks of different entities which have participated in the publication's creation (i.e. authors, venues, and affiliations). In this chapter, we focus on publication content in addition to citations and address the following research question:

RQ 4: *How can we use publication content to create new*

methods for assessing the quality of research publications?

More specifically, the goal of this chapter is to discuss how publication content can be utilised to produce new methods for assessing the characteristics of research publications, which would provide more meaningful information about research publication quality that the currently used metrics. Within this area, we propose *semantometrics* as a new class of research metrics which utilise text, and two novel methods, which are based on the idea of utilising semantic similarity of publications to identify bridges or brokers in the scholarly communication network; we also experimentally demonstrate the feasibility of calculating these methods. By designing new methods based on publication content, this chapter contributes to fulfilling Goal 1:

> Goal 1: *Design new methods for assessing the value research publications and evaluate these methods in comparison with existing research evaluation metrics.*

The first method aims at assessing the amount of a publication's contribution to the research field and is based on calculating semantic similarity of publications citing and cited by a given publication. In our method, each publication is viewed as a "bridge" between existing knowledge (the cited publications) and new knowledge developed using the publication (the citing publications). A publication has a higher contribution if it creates a "long bridge", e.g. by pushing its field further forward, or by bridging more distant areas of science.

The second method aims at characterising types of research collaboration to provide an early indication of potential future impacts and is based on semantic similarity of authors (represented by their publication record) who participated in the publication's creation and on the authors' previous collaboration record. Our method is therefore focused

on two dimensions of research collaboration: collaboration frequency and inter-disciplinarity.

In this chapter we formally introduce these two methods and experimentally demonstrate the feasibility of their calculation. The content of this chapter is organised as follows. In Section 6.1 we introduce semantometrics as a new class of metrics for evaluating research. In Sections 6.2 and 6.3 we introduce our content-based methods for assessing a research publication's contribution and for categorising the types of research collaboration. We summarise our findings and conclude the chapter in Section 6.4.

6.1 Semantometrics

In Chapter 2, we have shown that over the recent years, there has been a growing interest in developing new scientometric measures that could go beyond the traditional citation-based bibliometric measures. This interest is motivated on one side by the wider availability or even emergence of new information evidencing research performance, such as article downloads, views, and twitter mentions, and on the other side by the limitations of citation-based metrics for evaluating research performance in practice. The existing types of quantitative research metrics, including bibliometrics, webometrics, and altmetrics, are commonly based on counting the number of interactions (such as citations, social media mentions, website links) in the scholarly communication network (Chapter 2). This common characteristic means that these metrics have some shared limitations. For example, they fail to capture the sentiment and the motives behind the citation or the online mention. We have discussed the limitations of the existing methods in Chapters 1 and 2.

In parallel to this work, the growing Open Access movement is mak-

ing it easier to freely access and analyse full texts of research articles on a massive scale, creating new opportunities for the development of research metrics. However, text has not received as much attention in research evaluation as other types of data, possibly because it was not until recently (due to various copyright restrictions) widely and openly available. We believe there are a number of advantages to utilising text for the creation of new metrics: (1) in contrast to 'external' evidence of publication utility provided by the interactions, the publication manuscript provides 'internal' evidence more directly related to various aspects of quality, such as rigour and contribution (Chapter 3), (2) the manuscript is a type of data available immediately upon publication, (3) text can be combined with the interactions to give the interactions added meaning, for example by utilising the text to detect sentiment of the interaction.

A number of research studies, which we have reviewed in detail in Chapter 2, have previously made use of text. The oldest works combining text analysis and research analysis have used text to analyse relationships between scientific disciplines [He, 1999] and between science and technology [Noyons and van Raan, 1994] or to improve clustering [Glenisson et al., 2005]. Text analysis has also been used in the context of predicting future citation counts [Yan et al., 2012, Whalen et al., 2015], to evaluate research proposals [Hörlesberger et al., 2013], to analyse the distribution and recurrence of citations within scientific documents [Hou et al., 2011, Bertin et al., 2013, Hu et al., 2015], and for citation sentiment analysis [Athar, 2011] and citation classification [Teufel et al., 2006, Valenzuela et al., 2015]. While a number of researchers have successfully made use of text for various related tasks, significantly less studies have focused specifically on developing new research evaluation methods which utilise text, and the existing studies applicable in this area have been largely limited to studying and classifying citation context. However, we believe

text analysis offers many more opportunities for improving the existing metrics and developing new metrics. To demonstrate this point, in the remainder of this chapter, we will present two new methods based on publication content, which can be used to create new research metrics.

Furthermore, we propose *semantometrics* (a compound of words *semantic* and *metrics*), a new class of metrics for evaluating research. In contrast to the existing types of metrics, such as bibliometrics, webometrics, and altmetrics, semantometrics are not based on counting the number of interactions in the scholarly communication network, but build on the premise that text is needed to assess the quality of a publication. In Chapter 3, we have studied the concept of research publication quality, and in Chapter 5, we have picked one of these aspects (research contribution) and studied the performance of the existing metrics in assessing this aspect of quality. We believe utilising the publication manuscript provides an opportunity for improving this performance and for creating new metrics able to capture different aspects of quality, such as rigour (Chapter 3, which the traditional metrics may struggle to capture. To demonstrate the possibilities that utilising text offers, we develop two new methods for assessing and analysing the value of research publications. We then empirically test both methods on real datasets and provide analysis of the results. These methods are presented in Sections 6.2 and 6.3. A summary of our findings and contributions is provided in Section 6.4.

6.2 A semantic similarity measure for assessing research publication's contribution

In this section we present a novel semantometric approach for assessing a publication's contribution to the research field which utilises publication full text. In Chapter 3, we have shown that research contribution is often seen as one of the most important aspects of research quality. We have seen that research contribution is typically broadly described as follows (Chapter 3):

> *A creative/intellectual advance that makes a contribution to the field and state-of-the-art (such as new paradigms, theories, ideas, interpretations, methods, findings, problems, forms of expression), distinctive, or transformative work.*

As this description is very broad, we focus on one part of this description – "contribution to the field and state-of-the-art". Specifically, we are interested in analysing how far the state-of-the-art was moved forward as a result of the publication in question. Although this is only one type of contribution a publication can provide (for example, this definition does not capture how the publication contributed to professional practice), we will further refer to this method and the resulting metric as *contribution*, but understand it to mean specifically how far the state-of-the-art was moved forward thanks to the publication in question.

To assess the amount of research contribution a publication generated, we view the publication as a "bridge" between the state-of-the-art (the publications referenced by the publication in question) and the future work created thanks to the publication (the publications citing the

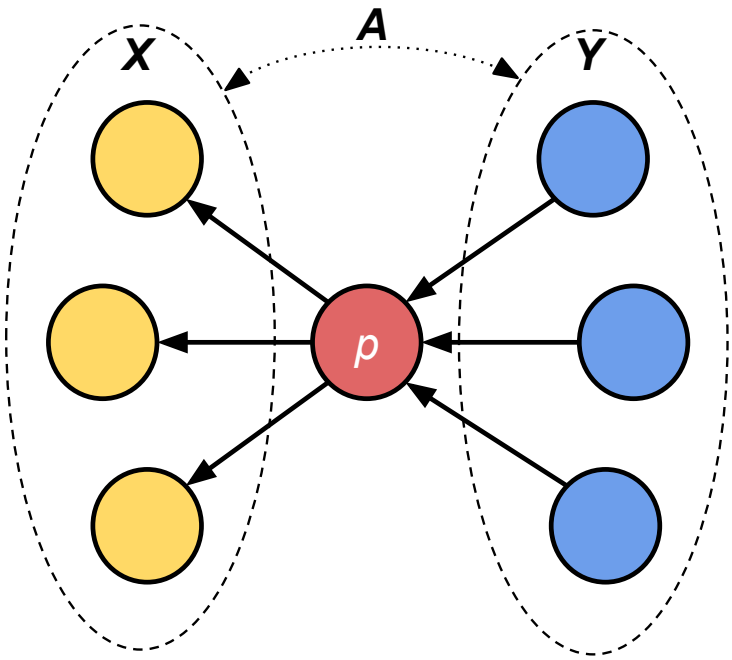

Figure 6.1: A visual depiction of the semantic distance (set of edges denoted as A) between the publications cited by publication P (set of yellow nodes denoted as X) and publications citing P (set of blue nodes denoted as Y).

publication in question), and we measure the semantic distance between these two sets of publications. This situation is depicted in Figure 6.1.

The intuition behind using the semantic distance between citing and cited publications is that while the cited papers are representative of the state-of-the-art in the domain of the publication in question (the publication itself contains only a fraction of the knowledge on which it is built, while the cited publications represent this knowledge more completely), the citing publications represent areas of application of the publication in question. Other research studies have used semantic distance between the publication and the cited publications to assess novelty [Yan et al., 2012] or the distance between the publication and publications that cite it to predict future citations [Whalen et al., 2015]. For measuring sci-

entific impact, both approaches suffer from some drawbacks. A metric based on the publication to cited distance would be easier to manipulate by careful selection of references. On the other hand, a metric based on the citing to publication distance disregards the amount of new information (originality/novelty) added by the publication in question. To overcome these issues, we have designed a metric which takes both the citing and the cited publications into account. The assumption is that useful innovation will propagate in the form of new knowledge to the citing publications, leading to a higher distance between the cited and citing publications. Based on the idea of measuring semantic distance between the citing and the cited publications, we create a metric that can be used to assess the amount of research contribution provided by a publication. This new metric is presented here in Section 6.2.1. Based on the definition of the contribution metric, in Section 6.2.2 we discuss the criteria required of a dataset used for calculating the metric. In Section 6.2.2 we present results of an experiment in which we calculate and analyse the metric. We summarise our findings in Section 6.2.4.

6.2.1 Contribution metric

As we have explained in the previous section, our hypothesis states that the added value of publication p can be estimated based on the semantic distance from the publications cited by p to the publications citing p. This hypothesis is based on the process of how research builds on the existing knowledge in order to create new knowledge on which others can build. A publication, which in this way creates a "bridge" between what we already know and something new which will people develop based on this knowledge, brings a contribution to science. A publication has a high contribution if it creates a "long bridge" between more distant areas of science, or more distant ideas within a field.

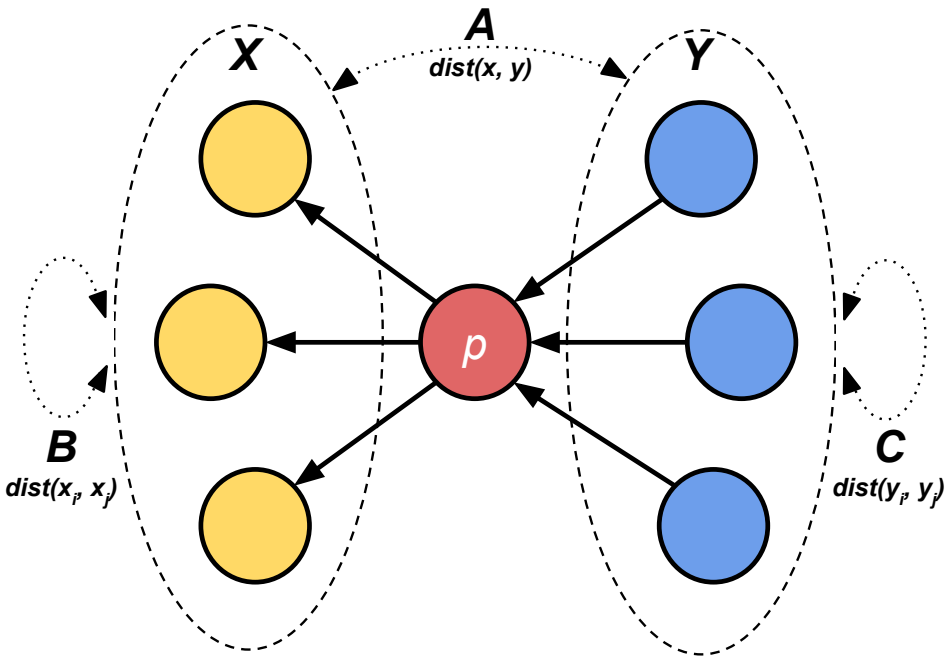

Figure 6.2: Explanation of $Contribution(p)$ calculation.

Building on these ideas, we have developed a formula assessing the publication's contribution, which is based on measuring the semantic distance between publications cited by p to the publications citing p:

$$contribution(p) = \frac{\overline{Y}}{\overline{X}} \cdot \frac{1}{|X| \cdot |Y|} \sum_{x \in X, y \in Y, x \neq y} dist(x, y) \qquad (6.1)$$

The distance $dist(x, y)$ is depicted in Figure 6.2. It can be seen that $contribution(p)$ calculation is based on calculating mean semantic distance between publications in the sets X and Y. In our initial definition and experiment, we utilise mean, however, in Chapter 7, we compare the performance of different statistics, such as median and range.

The numerator $\overline{Y}$ and denominator $\overline{X}$ in the first fraction of the

formula are both calculated according to the following equation:

$$\overline{X} = \begin{cases} 1, & \text{if } |X| = 1 \text{ or } |Y| = 1. \\ \frac{1}{|X|(|X|-1)} \sum\limits_{\substack{x_i, x_j \in X \\ x_i \neq x_j}} dist(x_i, x_j), & \text{if } |X| > 1 \text{ and } |Y| > 1. \end{cases} \tag{6.2}$$

The numerator $\overline{Y}$ is calculated in the same way. It can be see that the $\overline{X}$ and $\overline{Y}$ calculation also uses mean distance, however, in this case the mean is computed over distances between publications in the set X and in the set Y. It is expected that the distance *dist* used in Equations 6.1 and 6.2 is estimated using a semantic similarity measure on the full text of the publications, such as cosine similarity on $tfidf$ document vectors. Because semantic distance is a symmetric relation, the calculation of mean distance used in Equation 6.2 be optimised by disregarding repeating pairs in the calculation, that is by selecting the publication pairs using combination rather than permutation. The number of pairs is then equal to $\binom{|X|}{2}$ instead of $|X| \cdot (|X| - 1)$.

The first fraction in the above equation is a normalisation factor, which is responsible for adjusting the contribution value to a particular domain and publication type. The underlying idea is that, for example, in the case of a survey paper, it is natural that publications within the set X and also within the set Y will be spread quite far from each other. However, this is not a sign of the paper's contribution, but rather a natural feature of a survey paper. On the other hand, we believe that if a paper uses ideas from a narrow field, but has an impact on a very large field, it is a sign of higher contribution. In both cases, the first fraction of Equation 6.1 appropriately adjusts the value of the metric.

In practical terms, our method for assessing the contribution of a paper means that a paper with high contribution value does not need to be extensively cited, however it needs to inspire a change in its domain or

even define a new domain. This can be manifested by the changes in the vocabulary which are the result of a specific publication. Consequently, a very active scholarly debate about a survey paper in a specific subject generating many citations may have a lower value than a paper developing a new strand of research. An important feature of this idea is that our method does not require as long delay for assessment as the widely used citation counts (typically decades) and can be therefore applied also to fairly young researchers. It is hard to manipulate, it respects that scientific communities have different sizes in different disciplines, it is not focused on the quantity of publications as the h-index, but rather on the qualitative aspects.

6.2.2 Finding an experimental dataset

In order to fully test our hypothesis, it is necessary to acquire a dataset which would meet the following criteria:

Availability of full text is a prerequisite for testing our hypothesis as the calculation of similarity requires this information. The full text could potentially be substituted with abstracts, however, for our initial experiment we have decided to utilise full text. This also enables us to better test the scalability of the metric.

Density of the citation network refers to the proportion of references and citation links for which we can find articles and access their full text. This requirements has proved to be hard to satisfy. In order to carry out a representative test of our hypothesis, it is necessary to ensure that our dataset contains a significant proportion of articles citing each publication as well as documents cited by the publications. If a mean number of references per publication is ≈ 40 [Abt and Garfield, 2002], then the complete set of publica-

tions needed for the assessment of contribution of one publication would consist of 80 publications (we can expect the mean number of received citations will be approximately the same as mean number of references). If we wanted to examine the contribution of 100 publications, we would need a set of $\approx$ 8000 articles. Obtaining such set is time consuming due to restrictions on machine access to publications and subscription access rights.

Multidisciplinarity is important due to the assumption that transferring knowledge forward to more distant areas is an indication of a publication's research contribution. As a consequence, the dataset to test our hypothesis needs to contain a significant proportion of articles cited by the publication under evaluation as well as articles referencing the publication, and primarily those from different subject areas.

In Chapter 4, we have reviewed the existing research publication datasets. Unfortunately, none of the datasets we have examined meets all three of the above criteria. Our original expectation was that we will be able to find a subset of publications satisfying all of our criteria within the Open Access domain. For this reason we have first used the CORE dataset, which provides access to research papers aggregated from Open Access repositories and journals. However, as many references and citations are still from subscription based sources (non Open Access content which CORE cannot legally aggregate), we found the citation network too sparse for the purposes of our evaluation. However, we believe the situation will soon improve due to the government mandates ratified in many countries worldwide requiring the publishing of publicly funded research through the Open Access route. Consequently, we have experimented with enlarging the dataset by automatically downloading missing

Open Access documents from the publishers' websites. Unfortunately, we have found this task to be very difficult to accomplish due to a wide range of restrictions imposed by publishers on machine access to (even Open Access) publications hosted in their systems. Of the datasets we have previously examined, the MAG provides the most complete citation network, which we have analysed in in Chapter 4. However, the MAG does not provide publication full texts or even abstracts.

6.2.3 Experiment

With no existing dataset suitable for our task, we have decided to create a new small dataset meeting all the above mentioned criteria. This dataset was created by manually selecting 10 seed publications from the CORE dataset with varying level of citations in Google Scholar. Articles cited by these publications and referencing these publications that were missing in CORE were downloaded manually and added to the dataset. Only documents for which we found a freely accessible online version were included. Publications which were not in English were removed from the data set as our similarity calculation technique was not developed to deal with multilinguality. Table 6.1 provides a list of the 10 publications with the number of downloaded English documents. In total we were able to download 62% of all documents found as direct neighbours of the seed documents in the citation network. After removing non-English articles, the set was reduced to 51% of the complete citation network. The whole process took 2 days and the resulting dataset contains 716 PDF documents in total.

Table 6.1: The dataset and the results of the experiment. The documents are ordered by their citation score. Column $|Y|$ shows the number of citations each publication received and column $|X|$ the number of references (these letters match the letters used in Figure 6.2). The numbers outside of brackets represent the number of documents in English which were successfully downloaded and processed, while the numbers in brackets represent the size of the full set (i.e. numbers we retrieved from Google Scholar, which include publications in languages other than English and publications which were behind a paywall). The last column shows the contribution score.

#	**Title**	**Authors**	**Year**	$\|Y\|$	$\|X\|$	$c(p)$
1	Open access and altmetrics: distinct but complementary	Mounce.	2013	5 (9)	6 (8)	0.4160
2	Innovation as a Nonlinear Process, the Scientometric Perspective, and the Specification of an "Innovation Opportunities Explorer"	Leydesdorff, Rotolo and de Nooy.	2012	7 (11)	52 (93)	0.3576
3	Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments	Li et. al.	2010	12 (20)	15 (31)	0.4874
4	The Triple Helix of university-industry-government relations	Leydesdorff.	2012	14 (27)	27 (72)	0.4026

#	Title	Authors	Year	$\|Y\|$	$\|X\|$	$c(p)$
5	Search engine user behaviour: How can users be guided to quality content?	Lewandowski.	2008	16 (30)	12 (21)	0.5117
6	Revisiting h measured on UK LIS and IR academics	Sanderson.	2008	25 (41)	8 (13)	0.4123
7	How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management	Rafols et. al.	2012	39 (71)	70 (128)	0.4309
8	Web impact factors and search engine coverage	Thelwall.	2000	53 (131)	3 (10)	0.5197
9	Web Science: An Interdisciplinary Approach to Understanding the Web	Hendler et. al.	2008	131 (258)	22 (32)	0.5058
10	The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update	Harnad et. al.	2004	172 (360)	17 (20)	0.5004
	Total			**474 (958)**	**232 (428)**	

We have processed these articles using the CORE software and have produced the contribution score for the seed documents. This has been done in two steps: (1) We extracted text from all PDFs using a text extraction library[1], (2) we calculated the contribution score using the co-

[1]Apache Tika, `http://tika.apache.org`

sine similarity measure on $tfidf$ term-document vectors [Manning et al., 2008] created from the full texts as means for calculating the contribution score. More precisely, the distance used in the contribution score was calculated as $dist(d_i, d_j) = 1 - sim(d_i, d_j)$, where $sim(d_i, d_j)$ is the cosine similarity of documents d_i and d_j (the $1 - sim(d_i, d_j)$ value is often referred to as *distance* although it is not a proper distance metric as it does not satisfy the triangle inequality property).

The results for each of the 10 documents can be found in Table 6.1. It is interesting to notice there are quite significant differences between the contribution score of publications with very similar citation scores. A closer analysis of the results has showed that our approach helps to effectively filter out self citations to similar work or more precisely gives little credit for them. Also, the publication with the highest citation score does not have the highest contribution score, in fact its contribution score is lower than that of a publication which is cited ten times less. We argue that this indicates that publications with a fairly low citation score can still provide a high contribution to science.

Figure 6.3 shows a comparison of the contribution score with citation score and with the number of references. The line in both plots shows a linear model fit. The plot shows that the contribution score slightly grows with the increasing number of citations. This is an expected behaviour, because the likelihood that a publications influence a number of topics and disciplines generally increases with the citation count, however they are not directly proportional. For instance, we can find publications in the dataset with a lower citation score and a relatively high contribution score. This shows that even publications with low citation score can provide a high contribution to research. On the other hand, with the increasing number of references the contribution score slightly decreases. This shows that increasing the number of referenced documents cannot

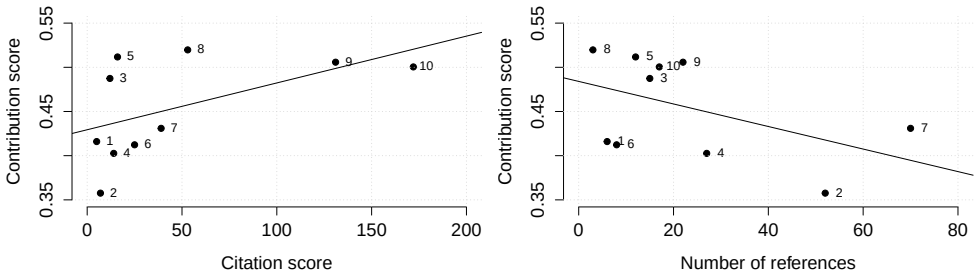

Figure 6.3: Comparison of the contribution score with citation score and with number of references.

be used to directly influence the contribution score.

6.2.4 Discussion and summary

The use of the current research publication metrics (such as bibliometrics, altmetrics, and webometrics) is based on a premise that the impact of a research paper can be assessed purely based on external data without considering the manuscript of the publication itself. We believe that utilising publication manuscripts provides many opportunities for developing new research metrics and improving the performance of the existing metrics. We have shown that new measures taking into account the manuscript of the publication can be developed. We believe that this idea offers a lot of potential for the study of this class of measures, which we call semantometrics. The results of our pilot study indicate that our measure based on semantic similarity of publications in the citation network is a promising method for assessing research contribution and should be further analysed on a larger dataset.

Furthermore, we have demonstrated the importance of developing datasets on which this class of measures can be tested and explained the challenges in developing them. The primary issue is the citation data sparsity problem, which is a natural consequence of publications referencing work from different disciplines and across databases. As systems

created by organisations that have bespoke arrangements with publishers, such as Google Scholar, do not share the data, there is a need for open data providers to join forces to create a single dataset spanning all scientific disciplines. Overall, we believe this situation demonstrates the need for supporting Open Access to research publications not only for humans to read, but also for machines to access.

6.3 Full-text based approach for analysing patterns of research collaboration

In the previous section, we have introduced the first semantometric measure for assessing the amount of research contribution a publication generated. The underlying idea behind the method is that each publication is perceived as a "bridge" between the state-of-the-art and the future work which made use of the publication, and the length of the bridge is assessed using semantic similarity methods on full text. In this section, we present a method based on a similar idea; however, in this case we focus on research collaboration. While the distance between citing and cited publications can aid in assessing research contribution, this information is not available for publications which have not been cited yet. To bridge this gap between a paper being published and it receiving the first citation, we have developed a method for analysing and categorising research collaboration, which does not depend on citation information.

Similar to our contribution metric, this method is based on identifying bridges or brokers in the scholarly communication network. It has been observed that in citation networks, bridging or cross-community citation patterns are characteristic for high impact papers [Shi et al., 2010]. This is likely due to the fact that such patterns have the potential for linking knowledge and people from different disciplines. The same holds true in

the case of collaboration networks, where it has been shown that newcomers in a group of collaborators can increase the impact of the group [Guimerà et al., 2005], and that high impact scientific production occurs when scientists create connections across otherwise disconnected communities from different knowledge domains [Lambiotte and Panzarasa, 2009].

However, the studies up to date have been predominantly focusing on analysing citation and collaboration networks without considering the content of the analysed publications. Our work focuses on analysing scholarly collaboration networks using semantic distance of the publications in order to gain insight into the characteristics of collaboration and communication within communities.

Our hypothesis states that the information about the semantic distance of the communities will allow us to better understand the importance and the types of collaboration. More specifically, in order to gain insight into the types of collaboration between authors, we investigate the possibility of utilising semantic distance in a coauthorship network together with the concept of *research endogamy* [Montolio et al., 2013]. In social sciences, endogamy is the practice or tendency of marrying within a social group. This concept can be transferred to research as collaboration with the same authors or collaboration among a group of authors. The concept of research endogamy has been previously used to evaluate conferences [Montolio et al., 2013, Silva et al., 2014] as well as journals and patents [Silva et al., 2014]. Research endogamy helps us to distinguish between groups of authors which are frequent collaborators from those which have not collaborated together frequently (and therefore potentially come from disparate or even disconnected communities). Author similarity then helps us to understand whether these collaborators come from a similar research background or whether they have

previously worked on dissimilar problems.

The content of this section is organised as follows. We start by presenting our research question (Section 6.3.1) and explaining the basic concepts used in our study (Section 6.3.2). In Section 6.3.3 we present the results of an experiment we conducted to analyse our method. A summary of our findings is presented in Section 6.3.4.

6.3.1 Research question

We investigate the relationship that exists between the tendency to collaborate within a group of authors and semantic distance of their respective research fields. In particular, we are interested in the distribution of the semantic distance of authors collaborating on a publication, the relation between the author distance and their endogamy value and whether, based on these two measures, there exists a typology of scientific collaboration across and inside of knowledge domains.

The rationale behind this approach is based on how research collaboration happens. In case the scientific collaboration spans across fields or disciplines, such research is likely to link the two disciplines and thus to provide opportunities for knowledge transfer, and for novel visions and ideas [Lambiotte and Panzarasa, 2009, Silva et al., 2014]. On the other hand, collaboration within one discipline can potentially increase the authors' performance [Lambiotte and Panzarasa, 2009].

We assume that based on the combination of semantic distance and research endogamy the types of research collaboration can be divided into four groups (Table 6.2). We believe this classification is a useful tool in characterising the types of research collaboration that goes beyond the traditional understanding of the concept of bridges as used in scholarly communication networks. While semantic distance allows distinguishing between inter- and intra-disciplinary collaboration, research endogamy

allows differentiating between emerging and established research collaborations.

	High endogamy	**Low endogamy**
High distance	Established interdisciplinary collaboration	Emerging interdisciplinary collaboration
Low distance	Expert group	Emerging expert collaboration

Table 6.2: Types of research collaboration based on semantic distance of authors, and their research endogamy.

The relation between author similarity and research endogamy is studied using the CORE dataset (Chapter 4). In this case we are able to utilise CORE, because to test our hypothesis we do not require a dense citation network, but instead, to be able to calculate author similarity and endogamy, need a sample of publications by each author. In this case, CORE is a perfect dataset, because it harvests whole institutional repositories, and will therefore contain a significant portion of publications of authors associated with that institution.

6.3.2 Basic concepts

This section introduces basic concepts used throughout this section. In particular it presents the definition of research endogamy and author distance used in our experiment.

Author distance

We propose to measure the semantic distance between authors of publication p as a mean semantic distance between all pairs of authors (Equation 6.3). Figure 6.4 illustrates which publications are used in the calculation.

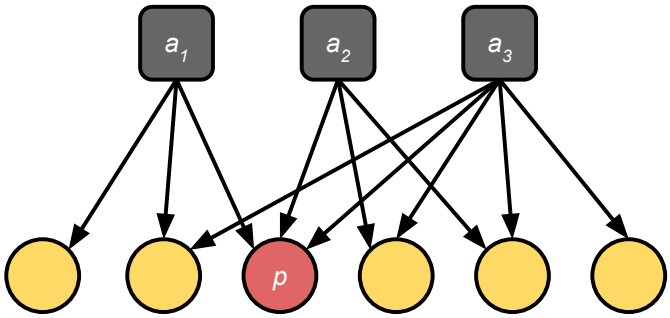

Figure 6.4: A sample network showing the set of publications (round nodes) and authors (squared nodes) used in the calculation of author distance and research endogamy of publication p.

$$a\_dist(p) = \frac{1}{|A(p)| \cdot (|A(p)| - 1)} \sum_{a_i \in A(p), a_j \in A(p), a_i \neq a_j} dist(a_i, a_j) \tag{6.3}$$

Here $A(p)$ is a set of authors of publication p. Similarly as in the case of the contribution metric, because semantic distance is a symmetric relation, this calculation can be optimised by disregarding repeating pairs in the calculation, that is by selecting the author pairs using combination rather than permutation. The number of pairs is then equal to $\binom{|A(p)|}{2}$ instead of $|A(p)| \cdot (|A(p)| - 1)$.

We calculate the distance for a pair of authors by concatenating the publications of each author into a single document. The distance of two authors is then calculated as semantic distance of two documents. While this is a very simplistic approach, it is also beneficial in terms of complexity of the calculation. Another approach would be to calculate the distance between every pair of publications of the two authors, perhaps omitting the publications they authored together. However, because the number of pair combinations of items of two sets has polynomial growth rate, this number would significantly grow in case of very productive authors. For this reason we chose to simplify the problem by adding all publications of one author into a single document.

Research endogamy

In order to distinguish between emerging, short-term and established research collaboration, we propose to combine the semantic distance with research endogamy value of the publication. The research endogamy of a publication is calculated based on research endogamy of a set of authors A, which is defined similarly as the Jaccard similarity coefficient [Montolio et al., 2013, Silva et al., 2014] (Equation 6.4). The authors and publications used in the calculation are depicted in Figure 6.4.

$$endo(A) = \frac{|d(A)|}{|\bigcup_{a \in A} d(\{a\})|} \tag{6.4}$$

Here $d(A)$ represents a set of papers written by all authors in A (each author in A has to be an author of each paper in $d(A)$). Higher endogamy value is related to more frequent collaboration between authors in A – a value of 1 means all authors in A have written all of their publications together. On the other hand, a group of authors who have never collaborated together will have an endogamy value of 0. Endogamy of a publication p is then defined as a mean of endogamy values of the power set of its authors [Montolio et al., 2013, Silva et al., 2014] (Equation 6.5).

$$endo(p) = \frac{\sum_{x \in L(p)} endo(x)}{|L(p)|} \tag{6.5}$$

Here $L(p)$ is the set of all subsets with at least two authors of p, $L(p) = \bigcup_{k=2}^{k=|A(p)|} L_k(p)$, where $L_k(p) = C(A(p), k)$ is the set of all subsets of $A(p)$ of length k.

Due to the way endogamy of a publication is currently defined, it has one significant limitation. Because the calculation of publication endogamy $endo(p)$ (Equation 6.5) is based on finding the power set of the set of publication authors, the number of times that author endogamy has be to calculated grows exponentially (this number is exactly

$2^{|A(p)|} - (|A(p)| + 1))$. That means that for a publication with 20 authors, author endogamy will have to be calculated for more than 1 million sets. However, it is not uncommon to have publications with more than a thousand authors, especially in some scientific disciplines. A potential simplification could be achieved by splitting the set $A(p)$ into groups of authors who ever collaborated together on any other publication than the reference publication p, and using these subsets for $endo(p)$ calculation instead of using the whole set $A(p)$. Because the reference publication p would not be considered in the calculation, this would potentially slightly lessen the resulting endogamy values. As the aim of this chapter is not redefining research endogamy, we used the existing equation, however we limited our dataset to publications with 25 and less authors.

6.3.3 Experiment

This section presents a basic overview of the dataset used in our experiment and the method used to obtain results. Furthermore, it provides a graphical representation of the distribution of research endogamy and author distance in the dataset and discusses the results.

Dataset

For this study, we have used a subset of the CORE dataset (Chapter 4 composed of:

- all full text documents which CORE harvested from Open Research Online[2] (ORO) repository (the Open University's repository of research publications),
- for calculating author distance and publication endogamy we also added all other full text publications found in CORE, which were

[2] `http://oro.open.ac.uk/`

authored by any of the authors of the publications harvested from ORO.

Table 6.3 presents overview statistics of the dataset. In the table the average number of collaborators is the mean number of different individuals an author collaborated with; and the total number of publications is the number of publications in the dataset after adding all other publications found in CORE, which were authored by any of the authors from ORO. More than 4,000 publications were analysed and the whole dataset included over 30,000 publications.

Fulltext articles from ORO	4207
Number of authors	8473
Average number of publications per author	7.61
Max number of publications per author	310
Average number of authors per publication	4.31
Max number of authors per publication	25
Average number of received citations	0.30
Average number of collaborators	80.23
Total number of publications	30484

Table 6.3: Statistics of the dataset used in our study of research collaboration.

We have selected the ORO repository as we needed a dataset containing the majority of publications of (at least a subset of) the academics. For this reason an institutional repository was a good candidate. We would like to note that we have not used any methods for disambiguating author names, as this problem is outside of the scope of this experiment.

Dataset processing

The following information was obtained from the CORE dataset:

- list of authors of each of the selected documents, and the publication record for each of these authors,
- number of times the publication was cited in CORE,
- fulltexts of the selected documents.

To calculate the author distance, we have used cosine similarity of $tfidf$ term-document vectors [Manning et al., 2008] created from the full texts. The documents were pre-processed by removing stop words, tokenising and stemming. Similarly as in the case of the contribution metric, the distance used in the author distance metric was then calculated as $dist(d_i, d_j) = 1 - sim(d_i, d_j)$, where $sim(d_i, d_j)$ is the cosine similarity of documents d_i and d_j.

Results

Figure 6.5 presents the distribution of both studied values, research endogamy and author distance. While the author distance is more similar to normal distribution, with mean 0.34 and standard deviation 0.19, the distribution of research endogamy is skewed with 50% of the publications having a value of less than 0.15. This is an interesting result, as it suggests it is not very common for authors to keep collaborating with the same academics.

Figure 6.6 shows a comparison of the two metrics with number of authors per publication. The lines in the plot represent a linear fit of the data. There is no correlation between author distance and the number of authors (Pearson $r = -0.09$). There is a very weak negative correlation between endogamy value and the number of authors (Pearson $r = -0.22$). This is an expected behaviour, because the likelihood that the endogamy value of a publication will be lower generally increases with the number of authors, however they are not directly proportional.

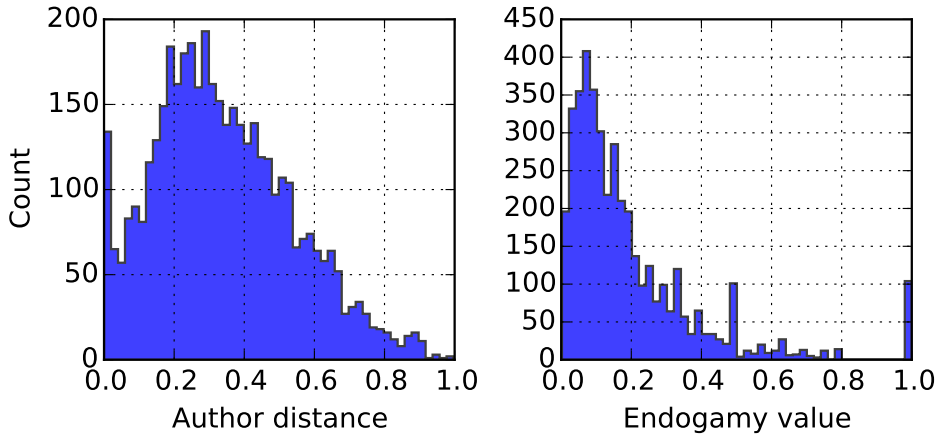

Figure 6.5: Distribution of endogamy value, author distance and number of citations.

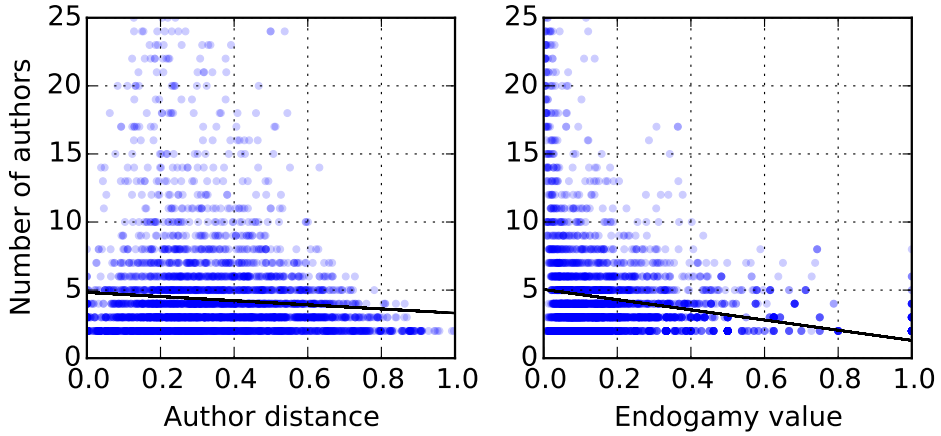

Figure 6.6: Author distance and endogamy value compared to the number of authors.

Figure 6.7 shows the relation between author distance and endogamy value. The lines in the plot represent the mean values of both data series. There seems to be one visible pattern in the data, which is the fact that very few publications fall in the category of high endogamy and high author distance, when using mean values as division lines. The proportion of publications which fall into this category is 0.07, while the

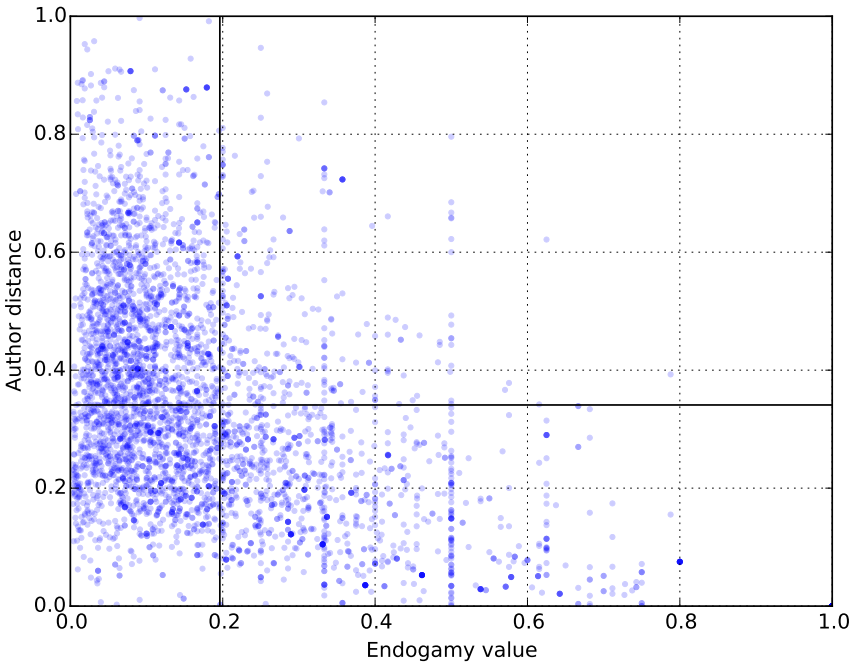

Figure 6.7: Author distance and endogamy value.

proportion of publications in the other categories varies between 0.27 and 0.38. This would suggest that collaboration across disciplines happens more often on a short-term basis. On the other hand, it seems that intra-disciplinary research does not tend to be done in one specific way, for example researchers do not tend to collaborate more often with the same colleagues.

We were interested whether certain types of publications attract more citations in general. Unfortunately the citation data was available only for a very small subset of publications. Figure 6.8 show the documents for which we had citation data (490 publications). The plot shows the relation between author distance and endogamy value, while the colour of the points indicates the number of received citations. The groups of publications with similar citation counts were selected based on percentile, the least cited group representing 50% of the publications while the

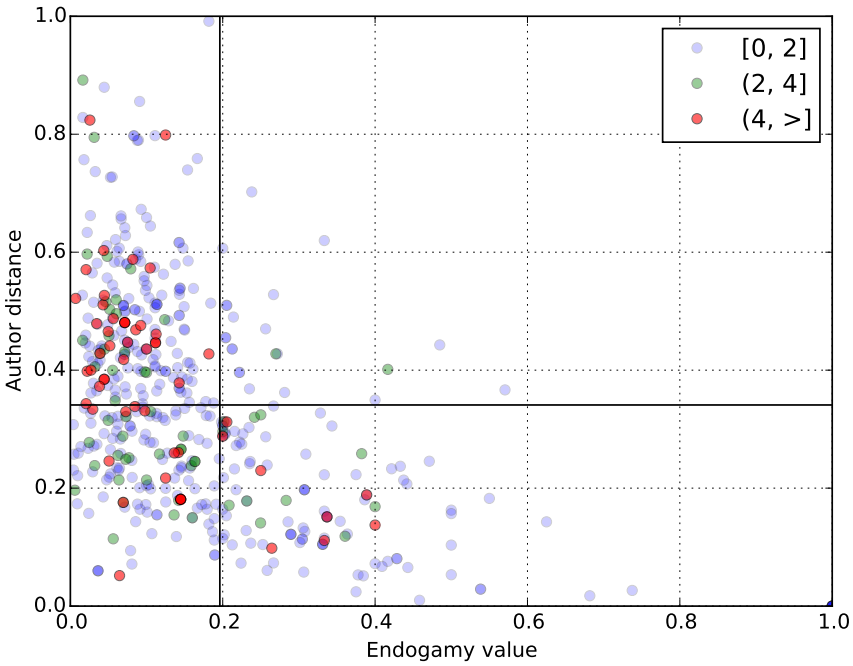

Figure 6.8: Author distance, endogamy value and number of citations.

highest cited group representing the top 10%. However, the differences between these groups are not large enough to be statistically significant. In our future work we would like to examine the relation between author distance, research endogamy and citation counts on a larger dataset.

6.3.4 Discussion and summary

In this section we have proposed to apply the semantometric idea of using full texts to recognise types of scholarly collaboration in research coauthorship networks. We have applied semantic distance combined with research endogamy to classify research collaboration into four broad classes and tested this classification using the CORE dataset. This classification can be useful in research evaluation studies and analytics, e.g., to identify emerging research collaborations or established expert groups.

While bridges have been the concern of many research studies, their identification has been limited to the structure of the interaction networks. In contrast to these approaches, our approach takes into account both the interaction network (coauthorship, citations) as well as the semantic distance between research papers or even communities when considering a group of authors which have not collaborated together frequently. This provides additional qualitative information about the collaboration, which has not been previously considered.

6.4 Conclusion

In this chapter, we have addressed the following question: "How can we use publication content to create new methods for assessing the quality of research publications?" Within this area, we have proposed *semantometrics* as a new class of research metrics which utilise publication content. Semantometrics are based on the premise that text is needed to assess the quality of a publication. To demonstrate the possibilities of semantometrics, we have designed two new content-based methods for analysing research publication quality and contribution. These two methods are based on the idea of utilising semantic similarity of publications to identify bridges in the scholarly communication network. We have developed the first method into a metric for assessing the amount of research contribution generated by a publication and the second method into a classification system for research collaboration. Furthermore, we have experimentally demonstrated the feasibility of calculating both metrics. By designing new methods which can be applied in research evaluation we have also contributed to Goal 1 which is focused on designing new methods for assessing the value of research publications.

Chapter 7

Evaluating research with semantometrics

> *I have learned to use the word impossible with the greatest caution.*
>
> – Wernher von Braun

In the previous chapter we have proposed *semantometrics*, a new class of metrics for evaluating research which utilise publication manuscripts. Furthermore, we have introduced two semantometric approaches for assessing research publications, which utilise semantic similarity of publications in the scholarly communication network to identify bridges or brokers in the network. The first method aims at assessing the amount of a publication's contribution to the research field, while the second method focuses on characterising types of research collaboration. In this chapter, we evaluate these methods for use in research assessment and address the following research question:

> RQ 5: *How can we interpret the performance of the content-based publication evaluation methods, and how do these methods compare to the existing metrics used in research evalu-*

ation?

We analyse and evaluate the proposed methods in two steps. We first perform a comparative analysis of the contribution metric with two other metrics: citation counts and Mendeley readership counts. This analysis is conducted on a dataset of over 300 thousand publications created by merging the Open Access dataset CORE, the citation network from MAG, and metadata from Mendeley (all three datasets are described in Chapter 4). The main goal of this analysis is not to advocate for the specific implementation of the contribution metric we have proposed in Chapter 6, but rather to analyse how the contribution metric behaves on a large dataset, and with respect to citation counts and Mendeley reader counts.

Next, we analyse both the contribution metric and the research collaboration classification method on our TrueImpactDataset (Chapter 4). The goal is to analyse the performance of our methods in distinguishing papers which provided very different amounts of research contribution and compare their performance to existing research metrics, particularly citation counts and Mendeley reader counts, among other metrics.

By comparing our methods to existing research metrics and evaluating our methods on a large collection of research publications, this chapter also contributes to fulfilling Goals 1 and 2:

> Goal 1: *Design new methods for assessing the value of research publications and evaluate these methods in comparison with existing research evaluation metrics.*

> Goal 2: *Show that the developed metrics can be deployed in large document collections to improve the analysis of published research.*

The content of this chapter is organised as follows. First, in Section 7.1 we present results of the comparative evaluation of our contribution measure with citation counts and Mendeley reader counts. In Section 7.2 we provide an evaluation of the performance of our semantometric methods and existing bibliometric and altmetric methods on our TrueImpact-Dataset (Chapter 4). We discuss our findings and conclude the chapter in Section 7.3.

7.1 Comparative evaluation of the contribution measure

This section reports on the analysis we carried out to investigate the properties of the semantometric contribution measure which we have introduced in Chapter 6. To investigate the contribution measure we utilise correlation analysis, which in scientometrics is a commonly used method for analysing new research metrics [Thelwall and Kousha, 2015a] and has been used for example in [Costas et al., 2015] and [Thelwall et al., 2013]. The main area of interest to us is the relationship between the contribution measure and citation counts. As Thelwall and Kousha [2015a] point out, "a positive correlation between a new indicator and citation counts is empirical evidence that the new indicator reflects something related to academic communication, rather than being purely spam or random, and the strength of the correlation can suggest the extent to which the two are similar." Our motivation for comparing the contribution measure to citation counts is also the prevalence of the use of citation counts in research evaluation. While utilising metrics based purely on citation counts has been subject to much debate (Chapter 1), these metrics remain among the best known and most widely adopted. The aim of this comparison is not to find a perfect correlation with citation counts, but

rather to demonstrate how the contribution measure behaves in relation to more familiar metrics.

This section is organised as follows. In Section 7.1.1 we describe how we collected the data needed for our study, and in Section 7.1.2 we present some summary statistics of the dataset and of the three measures being compared (citation counts, Mendeley reader counts, and semantometric contribution). In Section 7.1.3 we present the analysis we conducted and discuss our results. We summarise our findings in Section 7.1.4.

7.1.1 Data collection

As we have explained in Chapter 6, to the best of our knowledge no dataset which would meet all of our criteria for an ideal dataset for semantometric research (availability of full text, dense citation network, multidisciplinary) currently exists. Therefore, to test the scalability of our method and provide an analysis on a large dataset, we chose to calculate our metric using publication abstracts. To create a suitable dataset, we have merged data from CORE, the MAG, and Mendeley (all three datasets are described in Chapter 4). For the purposes of the analysis, we needed access to publications, their citation counts, and the textual data (abstracts) of research papers citing or cited by these publications. While the CORE dataset contains research publications with complete metadata, the MAG provided us with citation data, and Mendeley provided us with additional metadata, including abstracts and usage data (Mendeley readership counts). To assemble this dataset, we did the following:

1. We took a sample of papers from CORE (those having a DOI). At the time of preparing the dataset there were over 3.3 million such documents in CORE.

2. Using the DOIs, we mapped these papers to MAG. Because not all DOIs were found in the MAG, this reduced the size of the dataset from 3.3 to 1.6 million documents.

3. From the MAG, we identified the DOIs of all papers that are cited by the CORE papers or that cite the CORE papers. This resulted in a dataset of about 12 million documents (including the 1.6 documents from CORE) connected by 44 million citations.

4. We used the DOIs of all 12 million documents to retrieve additional metadata, such as readership counts, titles, and abstracts using the Mendeley API.

Using this procedure, we created a dataset containing information about 1.6 million papers from CORE, which included Mendeley reader counts and citation counts. The dataset also contained abstracts of over 10 million publications which cite or are cited by the papers from CORE. The Mendeley reader counts, which represent the number of Mendeley users having a particular article in their library, is a useful indicator in this context as it is an example of an alternative (usage-based) metric. Mendeley readership has been previously shown to exhibit a moderately strong correlation with citation counts [Li and Thelwall, 2012, Schlögl et al., 2014].

This dataset enabled us to calculate the contribution measure for 376,731 CORE papers. There are several reasons for this drop in numbers. First, as we have shown in Chapter 4, the MAG does not contain citation information for all publications. In fact, more than half of the publications in the MAG are disconnected from the citation network (i.e. have no references, and no citations). This reflects in the number of publications for which we were able to calculate the contribution measure, as for the calculation we need information about the papers a

publication cites and the papers the publication is cited by. Furthermore, a requirement for the calculation is having at least one citing and one cited publication with abstract for each of the CORE publications. We have used Mendeley to retrieve the abstracts of the citing and the cited publications; however, the Mendeley metadata is not always complete and the abstract is sometimes missing. Finally, a significant portion of all publications is never cited at all. In fact, according to some researchers, between 55 [Hamilton, 1991] and 90 percent [Meho, 2007] of research remains uncited, while a small proportion of publications receive high number of citations [Seglen, 1992]. While the first two reasons are effects of the databases we use to collect data, the latter reason is a natural effect of the citation distribution. As a consequence of these three effects, our dataset is reduced to 376,731 publications.

7.1.2 Dataset statistics

Table 7.1 presents overview statistics for our dataset.

Table 7.1: Dataset statistics. The numbers shown in this table include only those articles for which we were able to calculate contribution.

CORE articles matched with MAG	376,731
Average number of received citations	36.32
Standard deviation	87.38
Max number of received citations	11,659
Average readership	26.60
Standard deviation	56.27
Max readership	13,165
Average contribution value	0.8930
Standard deviation	0.0806
Max contribution	0.9999

Figures 7.1, 7.2, and 7.3 presented in this section provide descriptive statistics of the dataset. The variables of interest here are (1) citation counts (as a basis of bibliometric measures), (2) Mendeley reader counts (as a representative of altmetric measures), and (3) contribution (as a representative of semantometric measures).

Each of the figures presented in this section was produced using the same publications, specifically the publications for which we could calculate contribution (publications for which we have at least one citing and one cited publication with abstract). As a consequence, publications with zero received citations are not present in the data used for producing these figures.

Figure 7.1 shows the histogram of article citation counts in the dataset. As expected, the citation distribution is a long tail (power law) distribution. This is consistent with existing studies [Clauset et al., 2009]. Similarly, the readership distribution (Figure 7.2) exhibits the same properties as the citation distribution.

To confirm our data are consistent with previous studies, we have investigated the correlation between citation counts and readership. We found that the two metrics are correlated with Pearson $r = 0.3870$ and Spearman $\rho = 0.3870$ (a plot showing the relation of the two metrics can be seen in Figure 7.3).

This correlation can also be seen in Figures 7.4 and 7.5 which compare the readership values with citation counts using averaging. The goal of these plots was to analyse whether higher citation counts are associated with higher Mendeley reader counts and vice versa. As we have explained in the introduction, correlations with citation counts are typically used in studies analysing new research metrics to demonstrate whether the new metrics are associated with academic communication. However, using only a correlation test might not always tell the complete picture.

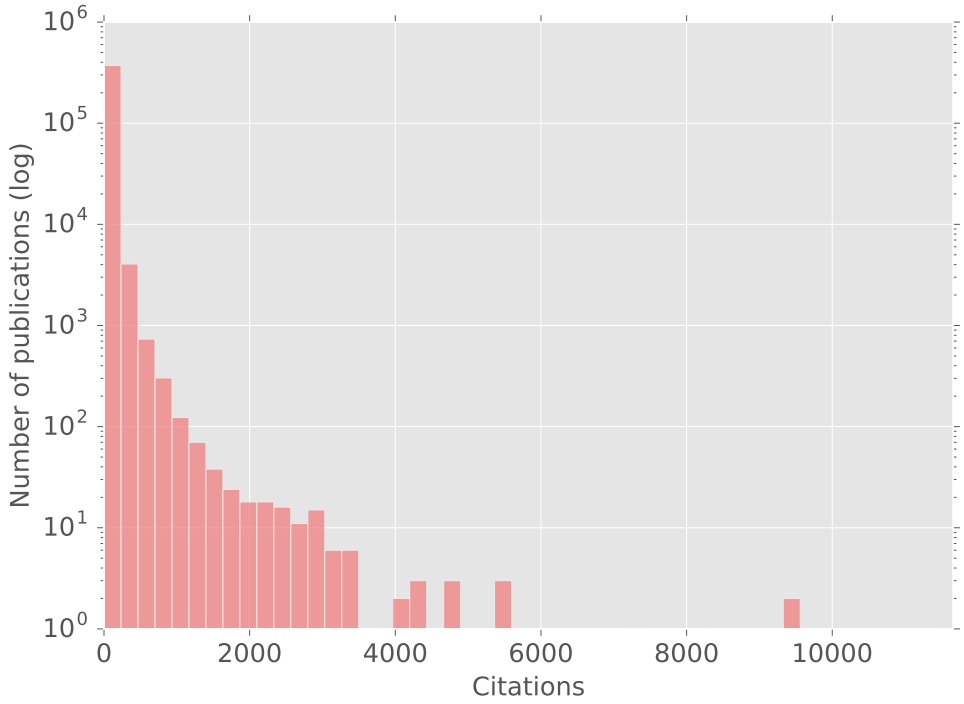

Figure 7.1: Histogram of publication citation counts.

For example, in the case of altmetrics and citation counts, the correlation may be low simply because altmetrics tend to be higher for newer articles (due to increasing uptake of altmetrics in general), whereas citation counts tend to be higher for older articles (due to having more time to accumulate citations) [Thelwall et al., 2013]. Therefore, to provide additional information about the relation between the two metrics, we devised the following method. To produce Figures 7.4 and 7.5, the data were split into 20 equally sized buckets. In case of Figure 7.4, we sorted and split the data by article citations and calculated the mean readership value for each of the buckets. In case of Figure 7.5, we sorted and split the data by readership and calculated the mean of article citations. In both figures, mean values are represented by the height of the bars, and the horizontal line represents mean value calculated across all buckets.

Some interesting observations can be made from Figures 7.4 and 7.5.

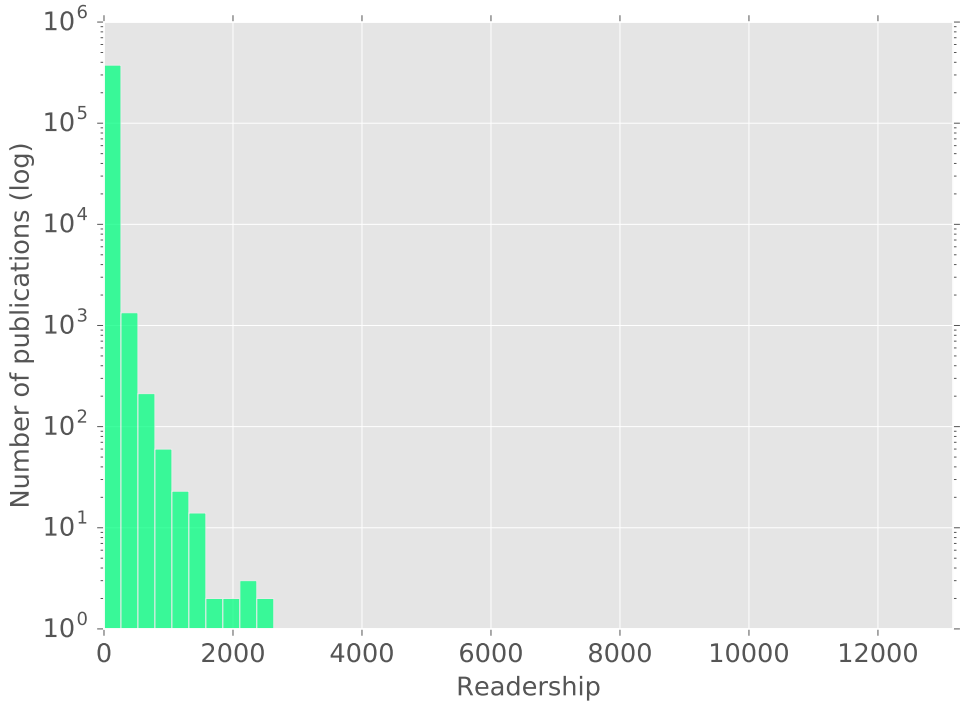

Figure 7.2: Histogram of publication readers counts.

Figure 7.4 there is a clear positive correlation between the averaged values. The figure shows higher citation counts are clearly associated with higher readership counts. However, the situation is slightly different in Figure 7.5. It can be seen that from a certain value of readership mean citation counts keep increasing. However, the figure shows publications with no Mendeley reader counts are cited higher than most publications with non-zero reader counts. This could be explained by the discrepancy between the "age" of citations and altmetrics which we have mentioned above. Older publications tend to have higher citations than newer publications simply because they had more time to attract citations; however, older publications also tend to have lower altmetrics because altmetrics did not exist before the creation of social media. These observations demonstrate our methodology reveals additional information which may be hidden in correlations and correlation plots such as the one in Fig-

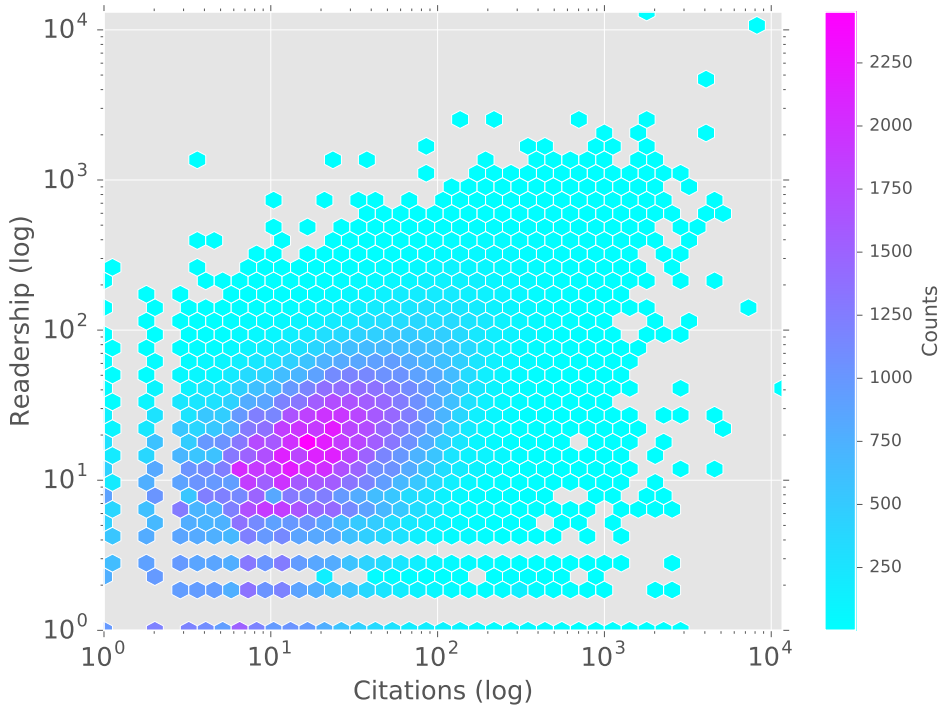

Figure 7.3: Relation between citation counts and reader counts.

ure 7.3.

As opposed to citation and readership distributions that resemble a power law, the contribution distribution (Figure 7.6) resembles a normal distribution. This has some implications. First, as a very large proportion of papers have no or just a few citations (and readers), it is difficult to evaluate these papers and compare the impact of these papers among themselves. Secondly, a power law distribution gives a skewed (and we believe incorrect) impression that the vast majority of research outputs are of poor quality. Finally, the fact that citation counts and contribution are distributed differently will likely reflect in lower correlation values between the two metrics.

One might argue that a normal distribution is a better reflection of the distribution of research outputs' quality. This is based on the assumption that the normal distribution is traditionally used to model

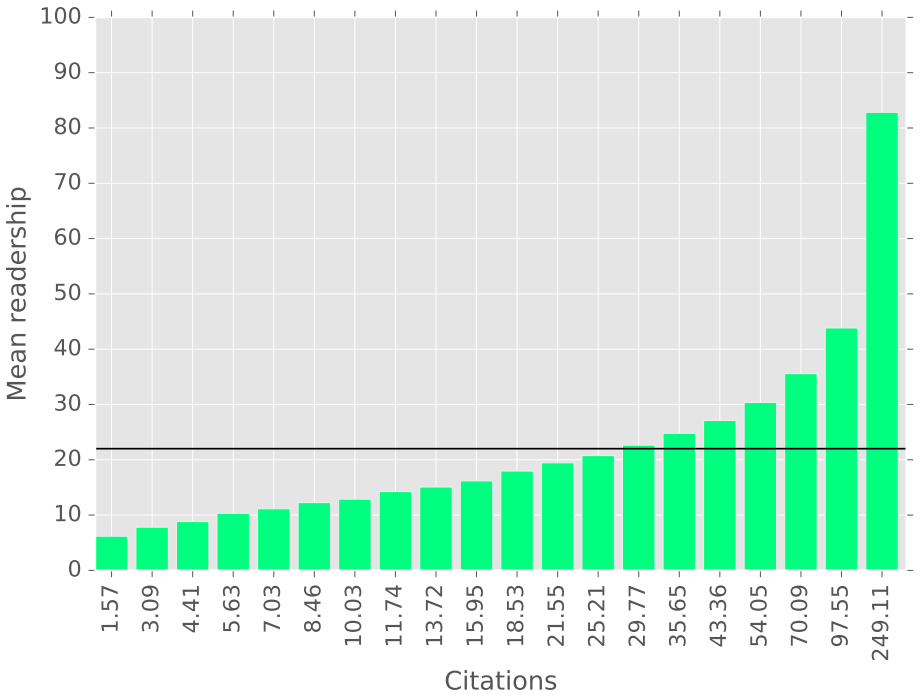

Figure 7.4: Comparison of citation counts with mean Mendeley reader counts.

the attainment of students, the performance of the workforce, and also the "quality" of papers as measured in the peer-review system. On the other hand, others might argue that the normal distribution is not a true representation of the papers' (particularly economic or societal) impact.

While the contribution distribution is skewed towards 1.0, we think this might be partly due to the fact that our contribution metric is calculated on article abstracts rather than full texts for the reasons of data availability. We assume that using full texts would result in a normal distribution with mean closer towards the centre of the graph.

Finally, the Pearson and Spearman correlation coefficient values for all three metrics are presented in Table 7.2. It can be seen there is a weak positive correlation between citation counts and contribution and a moderate positive correlation between citation counts and Mendeley

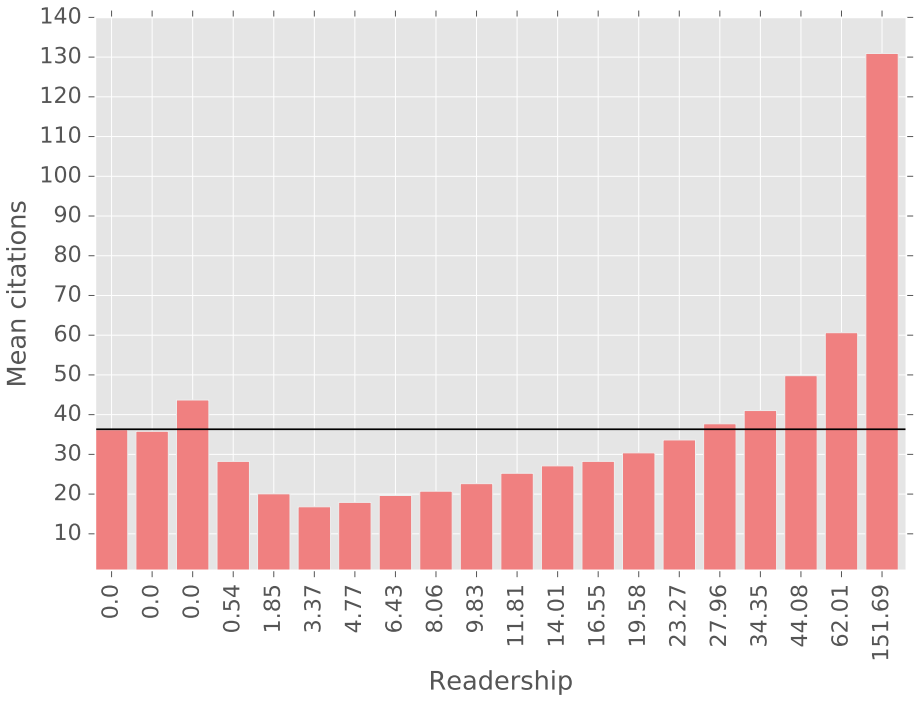

Figure 7.5: Comparison of Mendeley reader counts with mean citation counts.

reader counts. The correlation between contribution and readership is very low.

Table 7.2: Pearson's r and Spearman's ρ correlations between contribution, citation counts, and Mendeley reader counts, $p \ll 0.01$ in all cases.

	Contribution	**Citations**	**Readership**
Contribution	-	$r = 0.0866$	$r = 0.0444$
	-	$\rho = 0.1150$	$\rho = 0.0364$
Citations	$r = 0.0866$	-	$r = 0.3870$
	$\rho = 0.1150$	-	$\rho = 0.3455$
Readership	$r = 0.0444$	$r = 0.3870$	-
	$\rho = 0.0364$	$\rho = 0.3455$	-

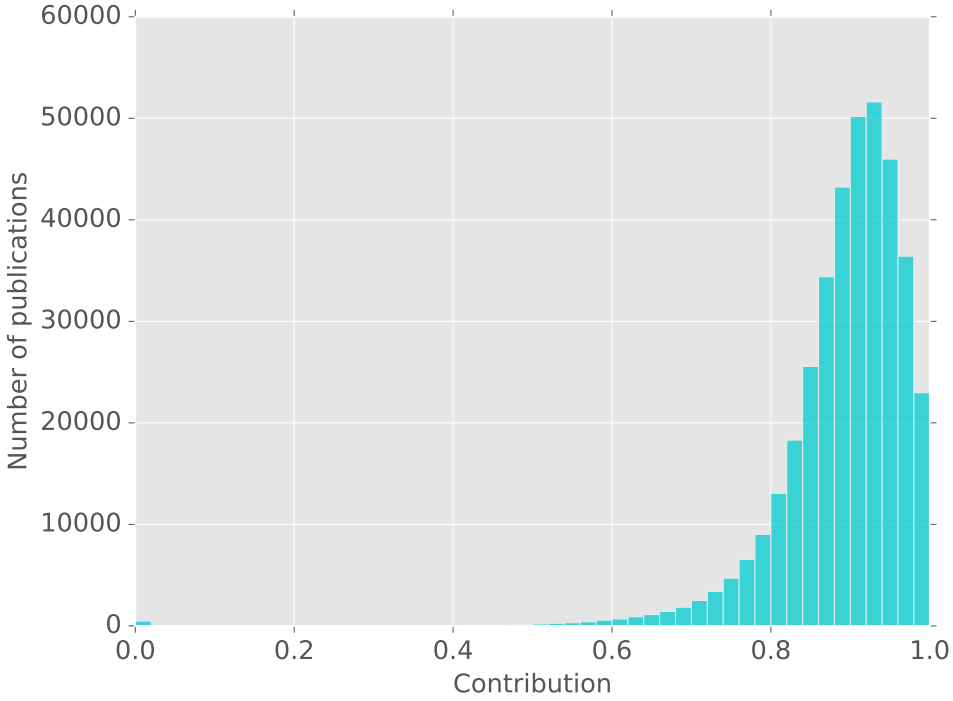

Figure 7.6: Histogram of publication contribution.

7.1.3 Analysis of the contribution metric

One of the main areas of interest to us is the relationship between citation counts and contribution, and between Mendeley reader counts and contribution. In contrast to citation counts and readership (which correlate with Pearson $r = 0.3870$ and Spearman $\rho = 0.3455$), we found a very low positive correlation between citation counts and contribution (Pearson $r = 0.0866$, Spearman $\rho = 0.1150$). We found no direct correlation between Mendeley reader counts and contribution (Pearson $r = 0.0444$, Spearman $\rho = 0.0364$). However, when we work with mean citation, contribution, and readership values, a clear behavioural trend emerges.

Figure 7.7 shows average contribution values compared with citation counts (to produce this figure we sorted and bucketed the publications by their citation value). The solid horizontal line represents the mean

value across all buckets which were split so that each contains the same number of data points. We can see that the behaviour of the contribution metric in relation to citation counts is not random. In fact, the averaged variables are correlated with Pearson $r = 0.7898$, $p \ll 0.01$ and Spearman $\rho = 0.6902$, $p \ll 0.01$. There is also low variance within the buckets, and standard deviation is 0.0390 across all buckets. However, when we calculate standard deviation on the whole dataset without bucketing, standard deviation is 0.0810. This confirms the consistency of the relation between the averaged values.

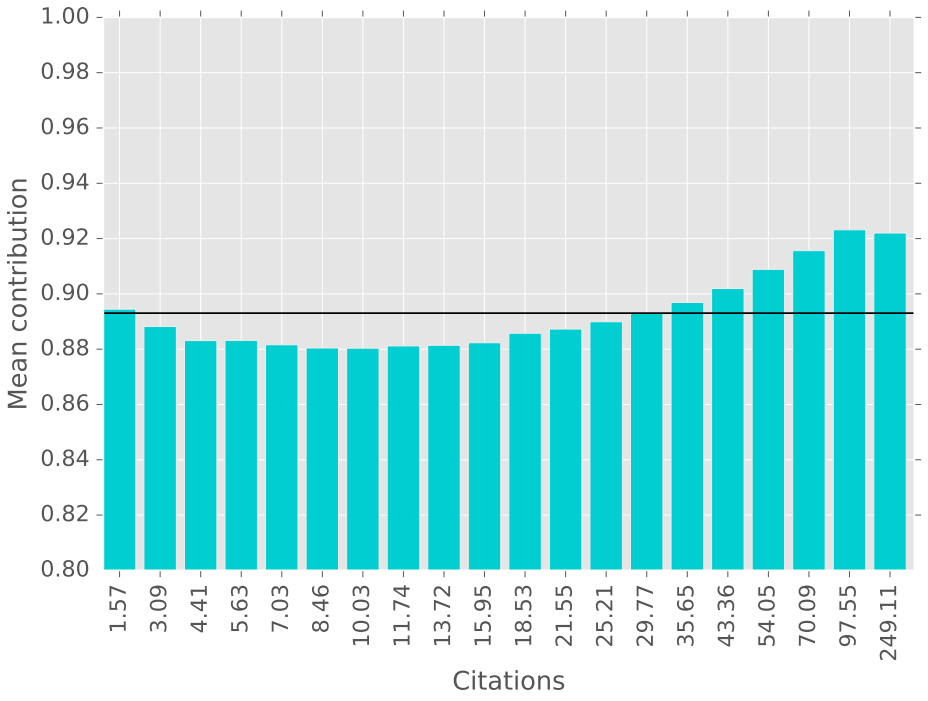

Figure 7.7: Mean contribution compared to citations.

We can observe that publications with a citation score above a certain threshold achieve on average consistently higher (above average) contribution (Figure 7.7). This threshold seems to be about 29 citations. It seems publications above this threshold are more likely to have higher contribution. However, once a paper receives around 90 citations, higher

citation counts do not lead on average to a higher contribution. One possible and highly simplified explanation for this could be that receiving around 90 citations is typically an indication of work with a major contribution (please note the opposite is not necessarily the case). Citation counts higher than 90 citations then typically reflect the size of the target audience community (visibility) rather than higher contribution of the underlying research work.

A similar observation can be made from Figure 7.8, which shows the relationship between mean citation count for a given area of contribution (to produce this figure we sorted and bucketed the publications by their contribution value). Note that each bucket contains the same number of data points. This relationship demonstrates Pearson correlation of $r = 0.9437$, $p \ll 0.01$ with the Spearman's $\rho = 0.8886$, $p \ll 0.01$, thus the relationship is statistically significant. In this case, standard deviation is 39.83 across all buckets. According to this graph, there are no differences in mean citation counts above a certain contribution value. This value seems to be about 0.89. It seems that publications can achieve high contribution regardless of how many times they are cited. On the other hand, publications with less than average contribution are also less likely to be cited.

Figures 7.9 and 7.10 report on a similar analysis of the relationship between readership and contribution. The relationship between average readership and contribution shows that papers with a higher readership are more likely to exhibit a higher contribution with Pearson's $r = 0.8479$, $p \ll 0.01$ and Spearman's $\rho = 0.8268$, $p \ll 0.01$. Variance within buckets is again low (standard deviation is 0.0401 across all buckets).

However, as we can see from Figure 7.10, articles with contribution between 0.91-0.93 are the most likely to receive the highest readership. Higher contribution is associated with an increase in readership until

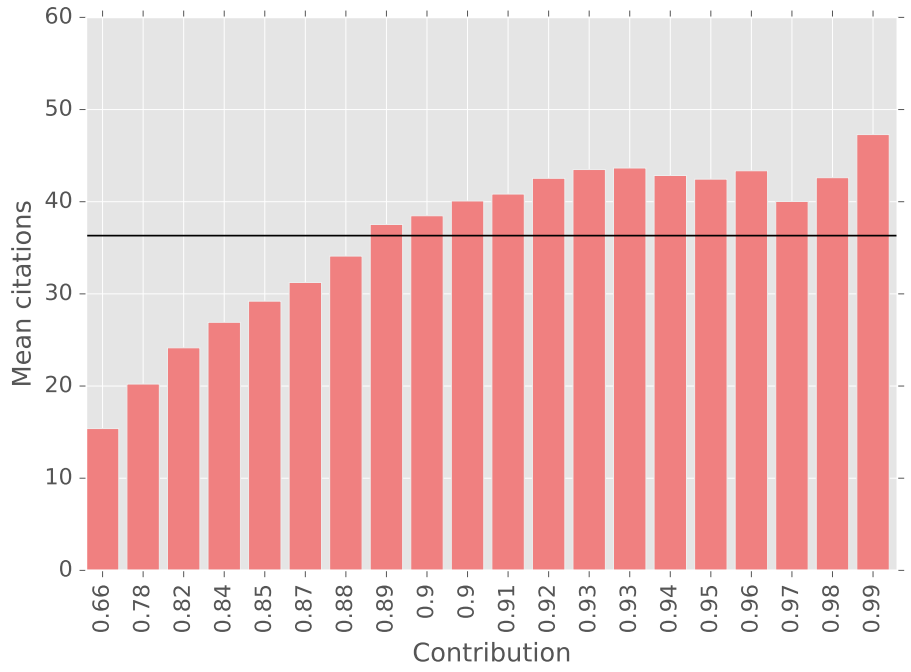

Figure 7.8: Mean citations compared to contribution.

it reaches 0.93 (Pearson's $r = 0.930$, $p \ll 0.01$, Spearman's $\rho = 0.9983$, $p \ll 0.01$). From then onwards, articles with higher values of contribution have on average lower numbers of readers (Pearson's $r = -0.9789$, $p \ll 0.01$, Spearman's $\rho = -0.9429$, $p \ll 0.01$ for the averaged values up till contribution of 0.93). Standard deviation across all buckets is 22.26. A possible explanation might be that as contribution is a measure of distance, papers that contribute to the emergence of new topics should be rewarded the most. However, such creation of a new discipline can be logically associated with the risk of a lower number of readers.

The Pearson correlation coefficient and Spearman's rank correlation coefficient of the averaged values for all three metrics are presented in Table 7.3.

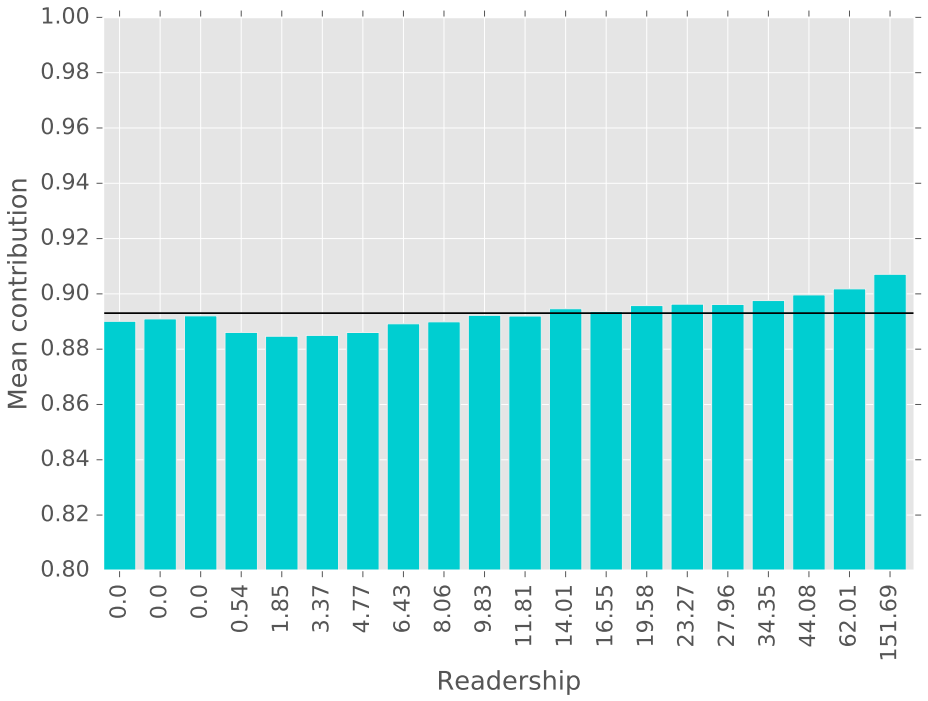

Figure 7.9: Mean contribution per readership value.

7.1.4 Summary

In this section we have provided a comparative analysis of our contribution measure with citation counts and Mendeley reader counts. This section therefore contributed to answering our research question "How can we interpret the performance of the content-based publication evaluation methods, and how do these methods compare to the existing metrics used in research evaluation?" To analyse our contribution measure and demonstrate it can be deployed in large document collections, we have conducted a comparative analysis of the measure with citation counts and Mendeley reader counts, which was conducted on a large collection of research publications created by merging data from CORE, the MAG and Mendeley.

This evaluation has revealed some interesting and useful properties

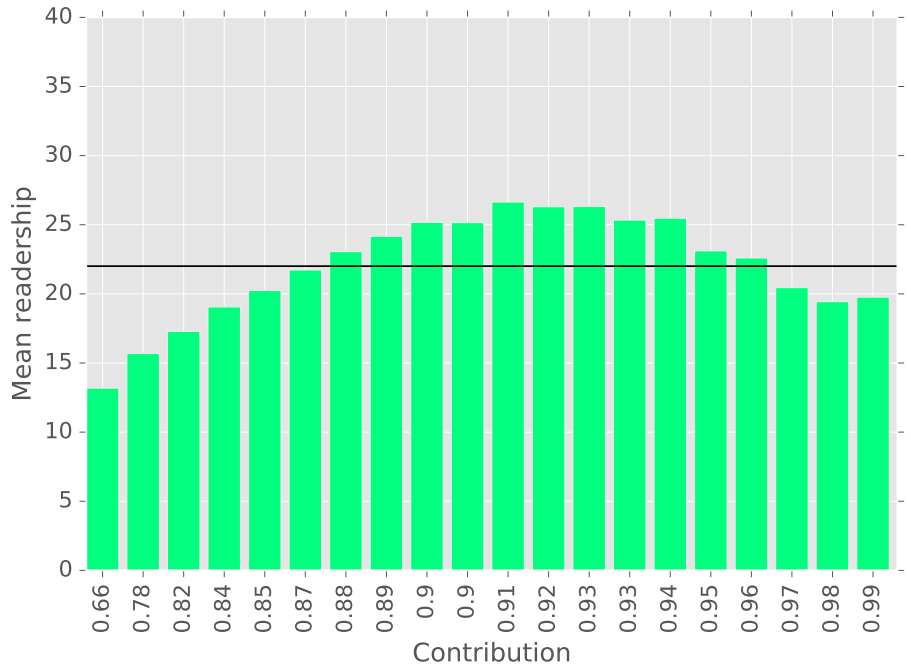

Figure 7.10: Mean readership per contribution value.

of the contribution measure. In particular, we have shown that contribution increases with increasing number of citations; however, after a certain threshold, higher citation counts do not lead on average to a higher contribution. One explanation for this is that receiving more than a certain number of citations reflects the size of the target audience (i.e. visibility of the publication) rather than higher contribution of the underlying research work. On the other hand, we have observed that there are no differences in mean citation counts above a certain contribution value (i.e. at first, citation counts increase with increasing contribution value, but stop increasing at a certain point). This suggests that publications can achieve high contribution regardless of how many times they are cited. We believe these are encouraging results consistent with our original intuition.

We would like to stress that, based on the results reported in this

Table 7.3: Values of Pearson's r and Spearman's ρ correlations between the averaged measures. In the table, the columns represent the variable used for bucketing (x-axis in the graphs) and the rows the correlated variable (y-axis). $p < 0.05$ in all cases.

	Contrib.	**Citations**	**Readers**
Contrib.	-	$r = 0.7898$	$r = 0.8479$
	-	$\rho = 0.6902$	$\rho = 0.8268$
Citations	$r = 0.9437$	-	$r = 0.9425$
	$\rho = 0.8886$	-	$\rho = 0.4849$
Readers	$r_{contrib.\leq 0.93} = 0.9300$	$r = 0.9868$	-
	$r_{contrib.>0.93} = -0.9789$		
	$\rho_{contrib.\leq 0.93} = 0.9983$	$\rho = 1.0$	-
	$\rho_{contrib.>0.93} = -0.9429$		

section, we are unable (nor is it our intention) to make claims regarding the superiority of the contribution metric in comparison to the existing metrics. Instead, this work presents an argument for studying the area of semantometrics more widely. Using the model example of the contribution metric, our goal is to encourage others to come up with new semantometric methods complementing (or going beyond) the contribution metric, to capture a variety of facets that good research publications exhibit.

7.2 Assessing research contribution with semantometrics

In the previous section we provided a comparative analysis of the contribution measure (Chapter 6) with citation counts and Mendeley reader counts. The goal of the analysis was to examine how the contribution

measure behaves on a large dataset, and with respect to citation counts and Mendeley reader counts. In this section we study the performance of both semantometric methods introduced in Chapter 6 (our methods for assessing contribution and for analysing collaboration) on our TrueImpactDataset (Chapter 4). In contrast to the previous study, the goal of this evaluation is to analyse the performance of the semantometric methods in distinguishing seminal publications from literature reviews and to compare their performance with other research evaluation metrics.

In Chapter 5 we have used the TrueImpactDataset to analyse the performance of two existing metrics, citation counts and Mendeley reader counts. We have shown that while citation counts distinguish between the two types of papers in the dataset with a degree of accuracy (63%, i.e. 10% over a random baseline), Mendeley reader counts do not work better than the baseline on this task (highest accuracy we achieve with Mendeley reader counts was 51.05%, while our baseline model achieved 52.87%). In this section we compare the performance of these two metrics (citation counts and Mendeley reader counts) with the semantometric methods we introduced in Chapter 6.

Furthermore, we present a detailed analysis of the semantometric measures. This part of the work is focused on examining whether our specific implementation of the both measures works well and whether there are any improvements we could make to improve their performance. For this part of the study we further develop the idea of analysing citation patterns in terms of content similarity. We extract a number of features describing content similarity of documents in a citation network and study how these features perform in our task.

The content of this section is organised as follows. First we describe our methodology and the different measures and features we compare in this study (Section 7.2.1). In Section 7.2.2 we describe the data sources

we used to collect data for the study. We report the results of our experiments in Section 7.2.3 and summarise our findings in Section 7.2.4.

7.2.1 Methodology

To study the performance of our semantometric methods we use our TrueImpactDataset (Chapter 4), a multidisciplinary dataset of research publications containing publications providing a very different amount of research contribution (specifically, the dataset contains seminal publications and literature reviews). Because we are interested in applying the results to research evaluation, our goal is to identify the most informative features which could be used in research evaluation methods. To be able to compare features in terms of performance, we approach this problem as a classification task.

We use the following methodology. For the publications in the dataset we collect and/or calculate three types of research measures: (1) semantometric measures (our contribution and collaboration measures), (2) bibliometric (citation-based) measures including citation counts and normalised citation counts, and (3) altmetric (web-based) measures including Mendeley reader counts. Furthermore, we extract a number of features describing semantic similarity of publications in a citation network, particularly a number of features which are used for the calculation of our contribution measure and other related features. In the context of this section, we will refer to both the research measures and the semantic similarity features simply as **features**.

Next, we compare the performance of all of the collected features in distinguishing seminal publications from literature reviews. In Chapter 3 we have shown that one of the most important factors influencing research publication quality is research contribution (i.e. how far was a field moved forward thanks to the publication). The approach we use in this section

therefore builds on the assumption that a good research evaluation metric should be able to distinguish publications that have changed a research field from those that have not. In Chapter 4 we have presented our TrueImpactDataset which we built for this task (i.e. analysing how well different methods assess research contribution). The dataset contains papers which are thought of as seminal (i.e. papers which inspired a change in their field) and papers whose aim is to provide a review of an area (i.e. papers which do not generate a change in their field). We study how well our features distinguish between these two types of papers. The following sections provide a complete list and description of all of the features we study.

Semantometric features

First, we calculate the semantometric contribution measure and a collaboration category (Chapter 6). The collaboration category is our only categorical feature and represents one of the four types of collaboration defined in Chapter 6, Table 6.2. As we are interested in studying the contribution and collaboration measures in more detail, we also collected a number of additional features.

Our contribution measure is based on calculating the semantic distance between the papers citing and the papers cited by a publication (these distances are labelled A in Figure 7.11). To provide a normalisation for cases such as when a publication references papers from a wide range of topics but influences a very narrow topic, the metric also includes a normalisation factor which is based on calculating semantic distance within the set of the cited papers (distances labelled D in Figure 7.11) and within the set of the citing papers (E).

There are two other types of links within a publications neighbourhood which interest us – links between the publication P and the papers

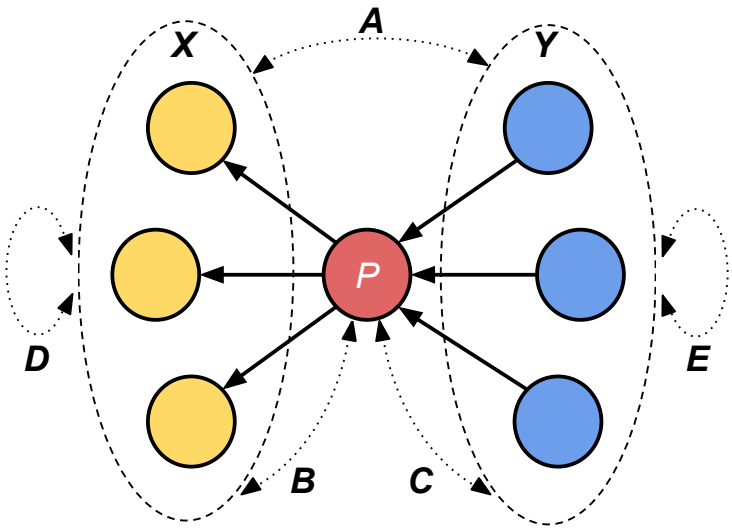

Figure 7.11: Neighbourhood of a single publication P and relations between publications in the neighbourhood which we investigate. The blue nodes (set Y) represent papers which cite the publication P and the yellow nodes (set X) represent papers which are cited by the publication P.

that cite it (these are labelled C in Figure 7.11), and links between the publication and the papers that it cited (links labelled B in Figure 7.11). Similarity between a publication and its references (i.e. relations labelled B) was previously used by Yan et al. [2012] to assess the publication's novelty. Recently, Whalen et al. [2015] have used similarity between a publication and the papers that cite it to predict future citation counts.

In this study we investigate all of the above mentioned relations (A-E) in terms of semantic distance between the publications participating in these relations. To measure the distance we use the cosine similarity measure of $tf - idf$ term-document vectors created from the publications' abstracts. We then calculate the distance of two publications as $dist(p_1, p_2) = 1 - sim(p_1, p_2)$, where $sim(p_1, p_2)$ is the cosine similarity between the $tf - idf$ weighted term vectors.

Each set of relations *A*-*E* described above is represented as a set of distances (for example a set of distances between a publication and each of its references). We define a set of metrics applied on the distributions induced by the distances. An example of the characteristics that we aim to distinguish is whether survey publications typically cite a wider range of topics than seminal publications and whether seminal publications tend to work within a narrower area. The metrics we use to describe the distance distributions, and which become our classification features, are (1) minimum, (2) maximum, (3) range (difference between maximum and minimum), (4) sum of the distances, (5) mean distance, (6) standard deviation, (7) variance of the distances, (8) 25^{th} percentile, (9) 50^{th} percentile (median), (10) 75^{th} percentile, (11) skewness (a measure of the asymmetry of the distribution, negative skew means the left tail is longer, positive skew means the right tail is longer), and (12) kurtosis (a measure of whether the data are heavy- or light-tailed, higher value means sharper peak). Because we describe each of the 5 distance distributions using 12 different metrics, we have 60 features (in addition to semantometric contribution and collaboration measures) describing each publication's neighbourhood.

Furthermore, the two features used to assign each publication a collaboration type (Chapter 6, Table 6.2) are added as separate features. These two separate features are *mean author distance* and *author endogamy*.

Bibliometric features

We extract four bibliometric features: (1) total number of citations per publication, (2) number of citations normalised by number of authors, (3) number of citations normalised by publication age, and (4) simplified relative citation ratio (S-RCR) as defined in Ribas et al. [2016] (Chapter 5).

The total number of citations per publication is probably the most frequently used method for the evaluation of research publications. Because older publications usually receive more citations simply because of more time available to collect citations, we add number of citations normalised by publication age as a feature. Furthermore, it has been observed that higher number of authors correlates with higher citation counts [Bornmann and Leydesdorff, 2015]. For this reason we also add number of citations normalised by the number of authors.

The S-RCR metric can be calculated as

$$S\text{-}RCR = \frac{ACR(p)}{(\frac{1}{|N_p|})\sum_{p' \in N_p} ACR(p')}, \tag{7.1}$$

where N_p is a set of publications which were co-cited together with publication p, and $ACR(p)$ is defined as

$$ACR(p) = \frac{citations(p)}{age(p) + 1} \tag{7.2}$$

The S-RCR metric is a simplification of the relative citation ratio (RCR) metric introduced by [Hutchins et al., 2016]. The idea behind a relative citation ratio is based on using a publication's co-citation network (Figure 7.12) to normalise the citation count of the publication. A co-citation network (nodes labelled N) of a publication P is defined as a collection of publications which appear in a reference list of any of the publications citing (the blue unlabelled nodes in figure 7.12) a given article. The underlying assumption is that articles which are cited together are similar in terms of a topic. A co-citation network therefore can be thought of as corresponding with the research area of the publication P [Hutchins et al., 2016]. This allows for accurate field- and time-normalisation of the citation count of publication P. Our motivation for including this metric in our analysis is that the S-RCR metric

was one of the winning solutions in the 2016 WSDM Cup Challenge (Chapter 5).

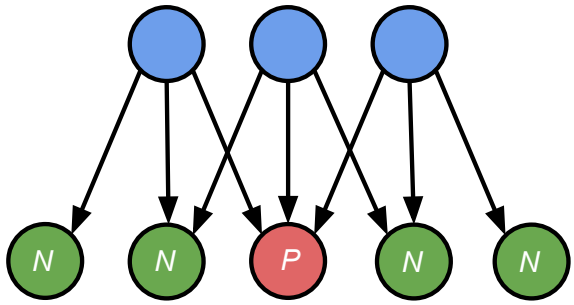

Figure 7.12: Sample co-citation (green nodes labelled N) network of a publication P.

Another frequently used citation-based method for evaluating the impact of scholarly publications is the Journal Impact Factor. However, we did not include any journal metrics in our study because not all of the publications in the dataset were published in a journal.

Altmetric features

We have collected three features related to a publication's social media visibility: (1) number of readers in Mendeley, (2) number of disciplines of the Mendeley readers, and (3) Altmetric score.

The first two features were collected from Mendeley. Mendeley was selected for our study because of its high coverage [Bornmann, 2015], and accessibility of the data. Mendeley provides information about how many people have bookmarked a certain publication, which we add to our dataset as a feature (i.e. total *reader count*). Mendeley also provides information about the research disciplines of the readers (there are 22 main research disciplines in Mendeley, e.g. "Biological Sciences", "Medicine", "Physics"). We use the information about the readers' disciplines as an estimation of size of the potential audience for the work presented

in the publication. This is our second altmetric feature, and represents the number of unique research disciplines of the readers of each publication and can therefore have a value between 0 (in case the publication has no readers) and 22 (total number of research disciplines provided by Mendeley).

Altmetric[1] is a service which collects and counts article mentions on social media (including Twitter, Facebook, and news sites) and aggregates the mentions into a single value (Altmetric score), which is a weighted sum of the different mention counts that Altmetric collects.

7.2.2 Data

Figure 7.13 shows a full neighborhood of one article (P) containing all components introduced above (citations, references and co-cited publications); the color-coding and labeling is consistent with Figures 7.11 and 7.12.

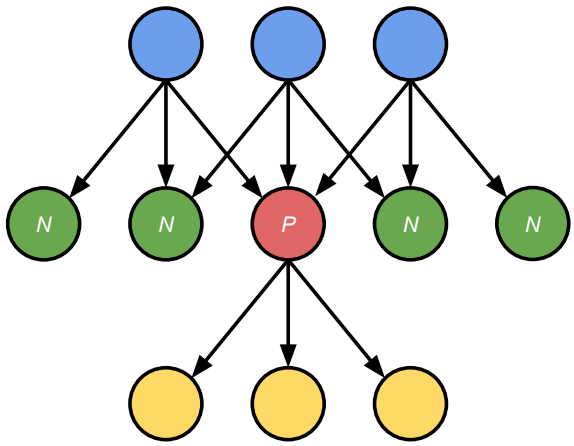

Figure 7.13: Full publication neighbourhood investigated in our study.

To gather all data needed to collect all features introduced in the previous section, we have used four data sources:

[1] https://www.altmetric.com/

1. TrueImpactDataset (Chapter 4), which is the core of our study and provides us with seminal publications and literature reviews,

2. Microsoft Academic Knowledge API[2] (Chapter 4) which we use to collect metadata (authors, year, venue, DOI, etc.) of the citing, cited, and co-cited publications (blue, yellow, and green nodes in Figure 7.13),

3. Mendeley API[3] (Chapter 4) which we use to collect abstracts (since Microsoft Academic does not contain abstracts) and information about readers,

4. Altmetric API[4] which we use to collect Altmetric score.

Table 7.4 shows for how many of the 314 publications found in the TrueImpactDataset we managed to get the needed metadata, and Table 7.5 shows how many additional publications we collected. The column "Total" in the second table was created by summing the totals for each paper, while the column "Unique" shows the number of papers of each type after removing duplicates (i.e. counting only unique publications among all references).

Table 7.4: Dataset size.

Publications in TrueImpactDataset	314
TrueImpactDataset publications in MA	298
Pubs with at least one citation in MA	269
Pubs with at least one reference in MA	215
At least one cit. and one ref. in MA	209

[2] `http://aka.ms/academicgraph`
[3] `http://dev.mendeley.com/`
[4] `https://api.altmetric.com/`

Table 7.5: Number of additional references we collected.

	Total	**Unique**
Number of citations	154,056	142,112
Number of references	13,599	12,562
Co-cited publications	4,999,682	2,269,364

Table 7.4 shows that in some cases the metadata we received from MA were not complete. We included in the experiment all publications for which we have at least one citing or one cited paper. We are therefore left with 269 core publications and 2,375,173 papers in total.

Features

For the 269 core publications we have collected 64 semantometric, 4 bibliometric, and 3 altmetric features. In case of bibliometrics and altmetrics we work with features representing the state-of-the-art in each area. In case of semantometrics our analysis is more exploratory.

The **semantometric features** we have collected are features **S1-S60** (Table 7.6) describing the *A-E* distance distributions (Figure 7.11), each of which is described using the following metrics: (1) minimum, (2) maximum, (3) range, (4) sum of the distances, (5) mean distance, (6) standard deviation, (7) variance of the distances, (8) 25^{th} percentile, (9) 50^{th} percentile (median), (10) 75^{th} percentile, (11) skewness, and (12) kurtosis. Furthermore, we have included feature **S61:** semantometric contribution, **S62:** semantometric collaboration category, **S63:** mean author distance, and **S64:** author endogamy.

The **bibliometric features** we have collected are **B1:** total number of citations per publication, **B2:** number of citations normalized by number of authors, **B3:** number of citations normalized by publication age, **B4:** simplified relative citation ratio (S-RCR).

Table 7.6: Features describing distance distributions *A*-*E*.

	A	**B**	**C**	**D**	**E**
min	S1	S13	S25	S37	S49
max	S2	S14	S26	S38	S50
range	S3	S15	S27	S39	S51
sum	S4	S16	S28	S40	S52
mean	S5	S17	S29	S41	S53
std	S6	S18	S30	S42	S54
variance	S7	S19	S31	S43	S55
p25	S8	S20	S32	S44	S56
p50	S9	S21	S33	S45	S57
p75	S10	S22	S34	S46	S58
skewness	S11	S23	S35	S47	S59
kurtosis	S12	S24	S36	S48	S60

The **altmetric features** we have collected are **A1:** number of readers in Mendeley, **A2:** number of unique disciplines of the readers in Mendeley, **A3:** altmetric score.

7.2.3 Experiments

We begin by comparing the properties of survey publications and literature reviews. We investigate how these two types of papers are situated with regard to the extracted features. To do this, we use the following methodology: we take all of the 269 core papers and for each of them collect all features defined in section 7.2.2. To understand which features might assist with the task we calculate an independent one-tailed t-test for each feature (except for the collaboration category feature which is categorical). The t-test is a measure commonly used to assess whether two sets of data are statistically different from each other. In other words, it helps to determine the features that can distinguish survey pa-

pers from seminal papers. To test the significance, we set the significance threshold at 0.05. Furthermore, for each feature we create a histogram and by comparing these histograms for the two publication types we gain insight into norms and placement of seminal and survey publications in terms of different research evaluation methods.

The complete results of the t-test are presented in Appendix D, Table D.1. Out of the 71 features, 32 result in p-value higher than 0.05. In this case we accept the null hypothesis of equal means. As the t-test tells us the values of these features are not significantly different for the two sets of papers, we remove these features from further analysis. The removed features describing the A-E distance distributions are crossed out in Table 7.7. Because we want to further study the performance of the bibliometric, altmetric, and semantometric measures, we do not remove these from the analysis even if their p-value is higher than 0.05. Specifically, features A1, A2, A3, S61, and S64 (reader count, number of unique readers' disciplines, Altmetric score, semantometric contribution, author endogamy) result in a p-value higher than 0.05 but are kept in our feature list.

From Table 7.7 it is obvious that there is not a single type of metric which describes all five distance distributions well. Furthermore, as most of the features describing the distribution E (distances among all citing papers) were removed, it seems this distribution does not offer much information for this particular task. This will be analysed in more detail.

Figures 7.14 and 7.15 show histograms of the remaining features, with seminal publications and literature reviews distinguished by colour. In both figures literature reviews are represented with dashed lines with circle points, while seminal publications with full lines with square points. The numbers in the legend of each plot show how many publications were used to produce each histogram (the numbers differ in case not all

Table 7.7: Removed and remaining features.

	A	**B**	**C**	**D**	**E**
min	~~S1~~	**S13**	~~S25~~	**S37**	**S49**
max	~~S2~~	**S14**	~~S26~~	~~S38~~	~~S50~~
range	~~S3~~	**S15**	~~S27~~	**S39**	**S51**
sum	~~S4~~	**S16**	**S28**	**S40**	**S52**
mean	**S5**	**S17**	**S29**	**S41**	~~S53~~
std	**S6**	**S18**	**S30**	**S42**	~~S54~~
variance	**S7**	**S19**	**S31**	**S43**	~~S55~~
p25	**S8**	**S20**	**S32**	**S44**	~~S56~~
p50	~~S9~~	**S21**	~~S33~~	~~S45~~	~~S57~~
p75	~~S10~~	~~S22~~	~~S34~~	~~S46~~	~~S58~~
skewness	**S11**	**S23**	~~S35~~	**S47**	~~S59~~
kurtosis	~~S12~~	~~S24~~	**S36**	**S48**	~~S60~~

publications had a value for a given feature). To preserve space we do not show here histograms of all of the remaining semantometric features S1-S60, but instead we select 15 features with interesting properties (Figure 7.15). Figure 7.14 shows the bibliometric, altmetric, and semantometric measures we are interested in in this study (features B1-B4, A1-A3, S61, S63 and S64).

In general, various metrics seem quite consistent across both groups. However, these metrics also reveal some important differences in citation patterns of seminal publications and literature reviews. First, one of our expectations is that useful innovation introduced by a publication will propagate in the form of new knowledge to the citing publications, leading to a higher distance between the publication and the citing publications (distance C) as well as between the references and citing publications (distance A). This is confirmed by higher average distances of both distributions in case of seminal publications (features S5 and S29 in

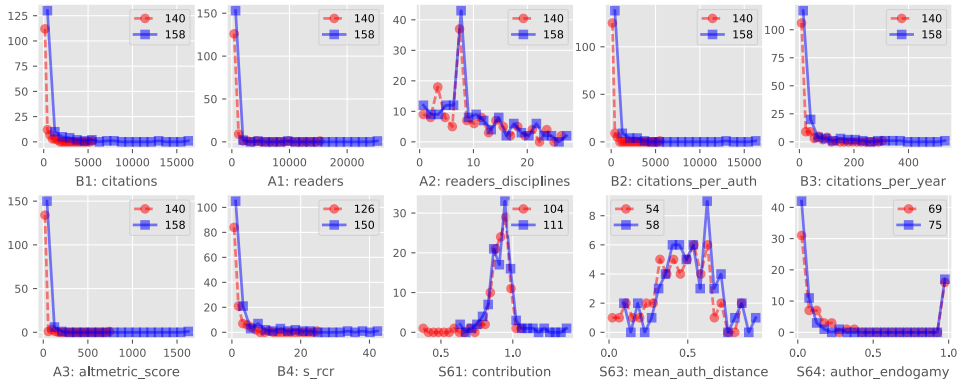

Figure 7.14: Histograms of the bibliometric, altmetric and semantometric measures.

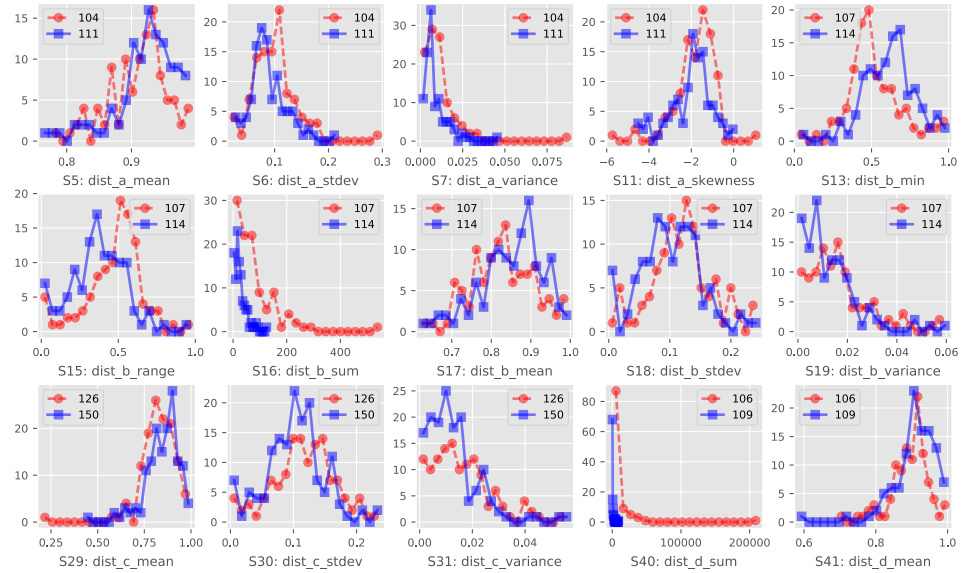

Figure 7.15: Histograms of selected features describing distance distributions *A*-*E* from Figure 7.11.

Figure 7.15). This is further supported by a lower standard deviation of the A and C distance distributions for seminal papers (features S6 and S30 in Figure 7.15).

Secondly, the distribution of distances between a publication and its references seems consistent with our expectations. In the case of lit-

erature reviews, the minimal distance between the publication and its references is on average smaller than for seminal papers (S13). At the same time, the difference between the most similar and most dissimilar reference is higher for literature reviews (S15). Even with the lower average distance between the literature review papers and their references (S17), the sum of distances between the publication and its references is higher for literature reviews than for seminal papers (S16), which is likely because reference lists of literature reviews are typically long. This feature could be used as a substitute for a simple reference count, which, although possibly a good indicator for distinguishing literature reviews and seminal publications, does not provide any useful information for assessing originality and research contribution, hence we remove this feature from further analysis. For the same reason we also remove feature S40 (sum of all distances among the references, i.e. sum of the D distribution).

The histograms of features describing the distance distribution E (features S49-S60) are very similar for both types of publications (except for S52). This was also confirmed by the t-test. Figure 7.16 shows all features describing distribution E. It seems the distances among the citing papers do not distinguish between seminal publications and literature reviews and therefore do not help in this task.

Finally, we analyse our semantometrics collaboration feature (S62). We calculate chi-square test, which is a statistical test for categorical variables for testing whether the means of two groups are the same, to test whether the seminal publications and literature reviews differ in terms of the semantometric collaboration category. The resulting p-value is 0.0218, which is lower than our significance threshold of 0.05. This tells us that the means of the two sets of papers differ. Figure 7.17 shows how the two classes are distributed with regard to the author distance and

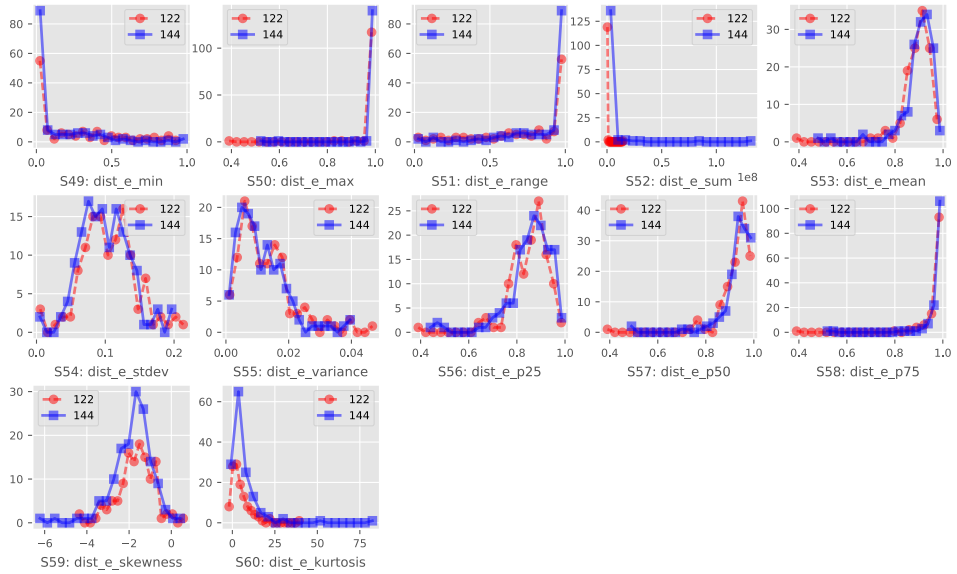

Figure 7.16: Histograms of features describing distances among citing papers.

author endogamy features. The horizontal and vertical lines in the figure represent mean values for each axis. The mean values are used to assign publications into the four categories. Furthermore, Figure 7.18 shows number of publications belonging to each collaboration category.

The figures show there are some differences between seminal publications and literature reviews. In particular, the main difference between the two classes is that emerging collaborations (i.e. when the authors have not collaborated frequently together previously) are in our dataset more common for seminal publications. On the other hand, literature reviews seem to be a result of established collaborations within a discipline. These observations are consistent with previous studies which have shown that cross-community citation and collaboration patters are characteristic for high impact scientific production [Shi et al., 2010, Guimerà et al., 2005, Lambiotte and Panzarasa, 2009]. We believe this is an encouraging result which suggest semantic distance of authors combined

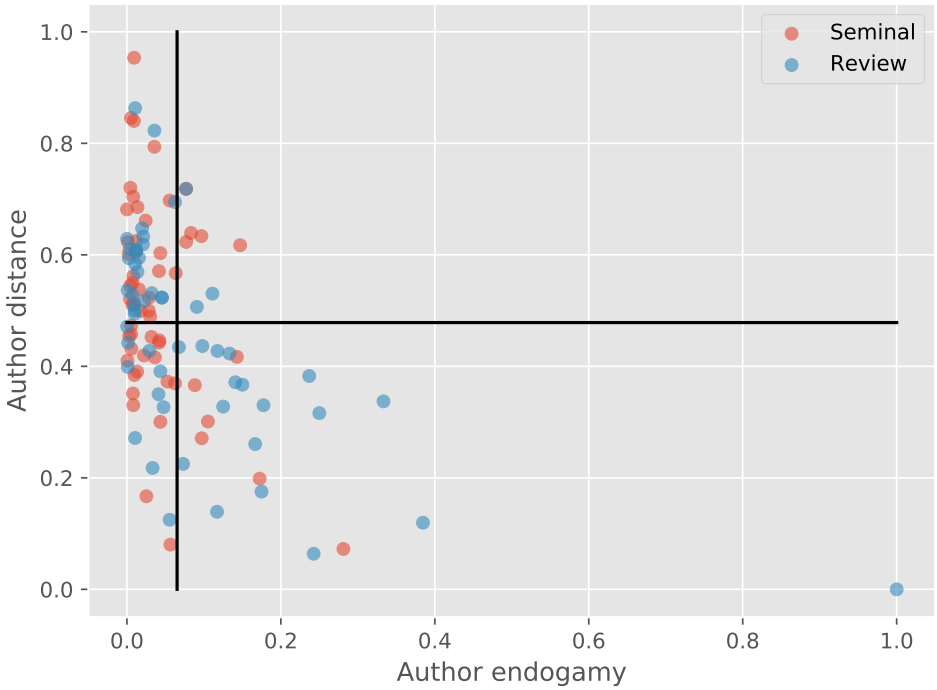

Figure 7.17: Distribution of publications according to author distance and author endogamy.

with their endogamy value might be helpful in providing early indication of future impacts of a publication.

Citation patterns and publication importance

In this section we explore the relation between the perceived importance of publications and the different metrics used to measure the importance. Although the above analysis of the separate features revealed distinct differences between the citation behaviour of seminal publications and literature reviews, we are interested in analysing whether the revealed patterns help in distinguishing important seminal publications from literature reviews better than current research evaluation methods. To do this, we approach this question as a classification task, which enables us to compare the features in terms of accuracy. We use four different meth-

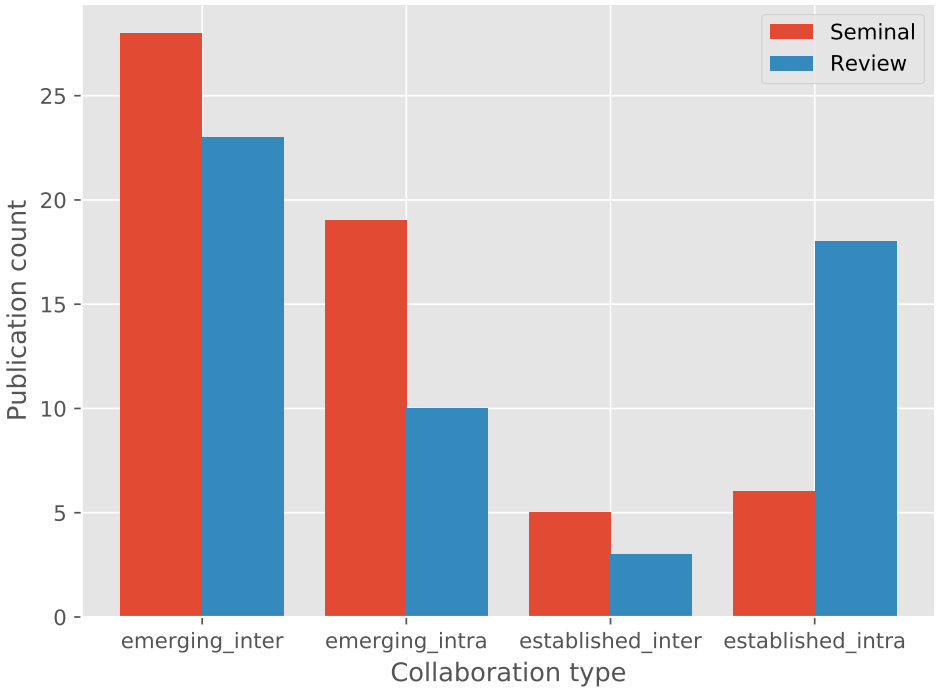

Figure 7.18: Number of publications belonging to each collaboration category across both publication types.

ods to identify important features and compare results obtained using the four methods.

In all classification experiments we use a leave-one-out cross-validation setup, that is we repeatedly train on all but one publication and then test the performance of the model on the publication we left out of the training. The performance is evaluated using accuracy, considering seminal papers as the positive class. All classifiers are compared against a baseline which always predicts the most frequent label. To produce Table 7.8 we train and test all classifiers twice. First we train the classifiers with all features selected in the previous section. However, as the author-related features (mean author distance and author endogamy) are only available for a small subset of the papers[5] (100 publications), in the

[5]This is because for some publications the data sources we used (Microsoft Aca-

second step we removed these two features. This increased the number of papers with complete data to 203. Table 7.8 shows classification accuracy using all of our selected classifiers. The results show that all classifiers outperform the baseline by up to ~23% (except for SVM which on the smaller dataset performs worse). We consider this an encouraging result, given the simple model and the fact we focus on very specific features.

Table 7.8: Classification accuracy using different classifiers.

	All features	**W/o auth. features**
# publications	100	203
Baseline	0.51	0.50
CART	0.70	0.67
Gradient Boosting	0.74	0.69
Gaussian Naïve Bayes	0.68	0.57
Support Vector Machine	0.40	0.54

To find feature importance we first use two models, Gaussian Naïve Bayes (GNB) and Support Vector Machine (SVM), to train each classifier using one feature at a time and rank features using classification accuracy obtained with each feature. This approach gives us a performance of each feature when used independently of other features. Table 7.9 shows the performance of the top 20 features for each model. The performance was obtained on all 203 publications (i.e. author features, distance and endogamy, were removed). The features are sorted in descending order of accuracy. Table 7.10 shows results obtained on the subset of publications which contain author information. Complete results are shown in Appendix D, Tables D.2 and D.3.

Tables 7.9 and 7.10 reveal some interesting results. First, it can be seen most metrics describing the B (distances between a publication and

demic and Mendeley) did not contain the data needed to calculate these features.

Table 7.9: Classification performance when using individual features and all 203 publications. The features are listed in descending order of accuracy, which is shown in brackets.

#	**GNB**	**SVM**
1	B range (0.65)	B min (0.66)
2	B min (0.65)	B range (0.65)
3	D min (0.61)	D range (0.64)
4	C variance (0.60)	D min (0.64)
5	D range (0.59)	D kurtosis (0.63)
6	C p25 (0.59)	D skewness (0.62)
7	D skewness (0.59)	Citations (0.60)
8	C stdev (0.58)	C sum (0.59)
9	D kurtosis (0.58)	B p50 (0.59)
10	D p25 (0.58)	E min (0.58)
11	E min (0.58)	B mean (0.58)
12	A variance (0.58)	B p25 (0.58)
13	A stdev (0.58)	E range (0.58)
14	B p50 (0.58)	S-RCR (0.57)
15	E range (0.58)	C p25 (0.57)
16	B mean (0.57)	Citations per year (0.57)
17	B p25 (0.57)	Altmetric score (0.55)
18	C mean (0.57)	E sum (0.55)
19	D mean (0.57)	A p25 (0.55)
20	Citations (0.56)	A skewness (0.55)

its references) and D distributions (distances among references of a publication) rank high. In fact, a single feature (*B range*) achieves accuracy which is almost as good as the best accuracy achieved using all features (Table 7.8). This suggests there is a referencing pattern which is different for seminal publications and literature reviews. However, the most interesting observation is related to distance distributions A (distances between publications citing a paper, and the references of the paper) and

Table 7.10: Classification performance when using individual features and the subset of publications which contains author information (100 publications). The features are listed in descending order of accuracy, which is shown in brackets.

#	**GNB**	**SVM**
1	B p25 (0.67)	B min (0.69)
2	B min (0.66)	B range (0.66)
3	D kurtosis (0.66)	D skewness (0.62)
4	B stdev (0.65)	D kurtosis (0.60)
5	B range (0.64)	B p25 (0.59)
6	D skewness (0.63)	D min (0.58)
7	B mean (0.63)	D range (0.58)
8	Author endogamy (0.61)	B p50 (0.57)
9	D mean (0.61)	S-RCR (0.57)
10	B variance (0.60)	A skewness (0.56)
11	A mean (0.60)	Author endogamy (0.55)
12	B p50 (0.59)	E min (0.54)
13	D variance (0.57)	D p25 (0.54)
14	D min (0.57)	Citations (0.53)
15	A p25 (0.57)	B mean (0.53)
16	D p25 (0.57)	E range (0.52)
17	E sum (0.56)	B variance (0.51)
18	Citations (0.56)	B max (0.51)
19	C stdev (0.56)	D variance (0.51)
20	S-RCR (0.56)	A stdev (0.51)

C (distances between a paper and the publications that cite it). Particularly when using Gaussian Naïve Bayes model, a number of metrics describing these distributions outperform citation counts. Namely, these metrics are mean and standard deviation. This confirms our observation from the previous section, and shows the distance between the citing and the cited publications works as a distinguishing feature between seminal

publications and literature reviews. Of the two author-related features (endogamy and author distance) endogamy achieved high accuracy with both models. This further confirms our findings from the previous section by showing that seminal publications may often be a result of new collaborations, whereas literature reviews are more frequently associated with established collaborations.

For a comparison we also use Gradient Boosting classifier (GBC) to rank features using feature importance learned by the classifier. We chose gradient boosting because we found it to work well on our dataset compared to other decision tree based classifiers we tested (classification and regression tree (CART) classifier, AdaBoost with CART as the base estimator, random forest classifier). Finally, we use recursive feature elimination (RFE) with a classification and regression tree (CART) classifier using information gain as a splitting criterion, and rank features in a reverse order of their elimination. Recursive feature elimination works by first training a classifier on all available features. In each step the least important feature is removed and the classifier is retrained on the remaining set of features. Tables 7.11 shows top 20 features for both classifiers and both setups (i.e. for publications with and without author features). In this case the features are not ranked independently, instead the rank is produced by training a classifier using all features at once.

Table 7.11 shows results obtained from the gradient boosting classifier and recursive feature elimination. The ranks produced these two methods are, especially for some features, quite different from the ranks produced when using independent features. We believe these differences show some features may not work as well when used independently, but provide useful information when combined with other features. For example, our contribution measure ranks high using both methods and in both setups. On the other hand, the top performing feature when used

Table 7.11: Feature importance obtained by training a gradient coosting classifier (GBC), and by recursive feature elimination (RFE). The features are listed in descending order of importance according to the two methods.

	RFE	**GBC**	**RFE**	**GBC**
#	203	203	100	100
0	C sum	D min	B min	D min
1	D min	Readers count	Contribution	D kurtosis
2	B min	Citations/auth.	D kurtosis	C skewness
3	D kurtosis	C skewness	C kurtosis	A stdev
4	C variance	C kurtosis	D min	Contribution
5	Contribution	Contribution	C variance	B min
6	C kurtosis	D kurtosis	B max	Author endo.
7	B p50	C sum	A mean	B range
8	A p25	E range	Author endo.	C p25
9	Readers count	B min	A variance	C kurtosis
10	E min	Citations/year	B stdev	C stdev
11	D skewness	B range	D stdev	D skewness
12	B stdev	S-RCR	Readers disc.	A skewness
13	A skewness	A stdev	D p25	Altmetric score
14	C skewness	D stdev	Auth. distance	S-RCR
15	A mean	E sum	Citations	Citations/auth.
16	B mean	D p25	B variance	D variance
17	D variance	D mean	D variance	Readers disc.
18	S-RCR	C stdev	S-RCR	Readers count
19	B range	B mean	Citations/year	C mean

independently (B range) ranks fairly low in terms of importance when used in combination with other features. Total citation counts also do not perform very well in this case.

7.2.4 Summary

In this section we have presented results of an analysis focused on evaluating our semantometric methods and comparing our methods with other research publication evaluation methods. The analysis was performed on our TrueImpactDataset (Chapter 4), and compared a number of metrics in terms of their performance in distinguishing seminal publications from literature reviews. Furthermore, we have studied citation patters of seminal publications and literature reviews in terms of semantic distance.

We have made a number of interesting observations. First, we were able to confirm that semantic distance between citing and cited publications is higher for seminal publications than for literature reviews. We believe this demonstrates studying citation patterns in terms of content similarity might provide more meaningful information which was not previously available. Furthermore, we observed different collaboration patters between seminal publications and literature reviews. While seminal publications in our dataset were more often a result of emerging collaboration, literature reviews were frequently associated with established collaborations within a discipline (rather than across disciplines). This suggests analysing research collaboration in terms of content similarity and collaboration frequency might offer an early indication of publication importance. Our contribution measure ranked high in terms of importance when used by models using a combination of features. Furthermore, we showed the underlying features used used in calculation of the contribution measure work in distinguishing seminal publications from literature reviews.

7.3 Conclusion

In this chapter we addressed the following research question: "How can we interpret the performance of the content-based publication evaluation methods, and how do these methods compare to the existing metrics used in research evaluation?" The evaluation was performed on two different datasets and using two different methods. First, we have conducted a correlation analysis of our contribution measure with citation counts and Mendeley reader counts. This evaluation has revealed some interesting and useful properties of the contribution measure. For example, we have observed that there are no differences in mean citation counts above a certain contribution value which suggests that publications can achieve high contribution regardless of how many times they are cited.

Next, we have evaluated our semantometric methods using our TrueImpactDataset, which enabled us to compare different measures in terms of classification accuracy. We have shown there are a number of features which describe citation patterns in terms of content similarity, which significantly outperform citation counts in distinguishing seminal publications and literature reviews on our dataset. This is the most important finding of this chapter as it demonstrates content analysis might provide additional valuable information for research evaluation.

There are a number of challenges in large-scale adoption of semantometrics. In our view, the two main challenges include:

Demonstrating the value of semantometrics: Despite their limitations, the use of bibliometrics in research evaluation is already deeply integrated into processes linked to recognition, such as grant funding and promotion. Introducing a new metric has limited chances of success due to many competing approaches being proposed in this area at this time. Convincingly demonstrating better

performance than widely used bibliometrics is therefore a necessary pre-requisite of success. This approach is very different from the axiomatic and ad-hoc approach in which the widely used measures, such as h-index, were established in the past. However, such performance demonstration is technically complicated to carry out. It cannot be simply achieved by demonstrating correlation with existing scientometric measures. In this area we face the challenge of developing datasets that can be used as the gold standard/ground truth for the evaluation of research metrics. While we have demonstrated a simple way in which such a dataset can be created, enlarging and broadening this effort is of paramount importance.

Large scale access to full text: Effective use of semantometrics requires unrestricted access to the manuscripts of research publications for text and data mining (TDM) purposes. In our study, we had to limit ourselves to the use of abstracts. At the moment, there doesn't seem to be any easy solution to this problem than to rely on a full text research papers aggregation systems (for Open Access content) and on the largely limited publisher TDM APIs (for toll access content).

Part III

Conclusion

Chapter 8

Conclusion

Somewhere, something incredible is waiting to be known.
– Carl Sagan

8.1 Introduction

Thesis we investigated new methods for assessing the value of research publications. To this end, the central research question studied in this thesis was:

> **How to effectively incorporate publication content into research evaluation to provide additional evidence of publication quality?**

The main motivation behind focusing on content was to show whether and how content can be exploited to develop research evaluation methods that are representative of research publication quality and to use this knowledge to improve the process of research evaluation. Multiple research questions arose from setting this objective, namely:

- Question 1: What is research publication quality and what factors influence it?

- Question 2: How can we evaluate the performance of metrics used in research evaluation for assessing the quality of research publications?
- Question 3: What is the relationship between the existing metrics used in research evaluation and the quality of publications?
- Question 4: How can we use publication content to create new methods for assessing the quality of research publications?
- Question 5: How can we interpret the performance of the content-based publication evaluation methods and how do these methods compare to the existing metrics used in research evaluation?

We started our research by investigating the concept of research publication quality. To discover the dimensions and aspects of the concept and to understand the importance of these aspects, we have carried out an in-depth literature review. To verify and expand the results of the literature review, we have conducted a survey which asked researchers about the importance of different aspects of publication quality. Our findings from this investigation have shown research quality is typically described in terms of three main criteria: originality (the contribution the publication/research provided), rigour (how well was the research performed and the publication written), and significance (what/who did the research/publication affect) (Chapter 3); our findings also provided an understanding of factors that influence each of these three dimensions of quality and how strongly they are related. These findings have inspired the way we have thought about the concept of research publication quality in the rest of the thesis, and they have been applied in the new research publication evaluation methods presented in Chapter 6.

Before developing new research metrics, it is necessary to understand how we can assess the performance of these metrics to understand

whether they work well and measure what was intended. We have addressed this question in Chapter 4. We investigated how research metrics are typically evaluated, and we built a new reference set complementary to the existing methods which can be used for validating research metrics.

Once we had a better understanding of the aspects of research quality and methods for evaluating research metrics, we focused on a selection of existing widely used metrics. In Chapter 5 we evaluated the performance of the selected research publication metrics using two different methods. First, we used our dataset developed in Chapter 4. This revealed that, while the existing metrics work on our dataset to a certain degree, there is room for improvement. Next, we have evaluated the performance of the existing metrics for ranking scholarly publications according to their importance. This evaluation was performed on a ground truth dataset of human judgement data. In this task we observed a similar performance as in the first evaluation. By combining information about publications and related entities (such as authors, venues, and affiliations), we were able to design a new publication ranking method with significantly better performance in this task. This is an important finding as it demonstrates improvements can be made to the existing research metrics to make these metrics more reliable.

Finally, we proposed *semantometrics*, a new class of research evaluation methods which utilise publication content. In contrast to the existing research metrics which rely on external evidence, semantometrics build on the premise that text is needed to assess the quality of a publication. To demonstrate the possibilities of semantometrics, in Chapters 6 and 7 we introduced and evaluated two semantometric methods. The key idea that these methods are based on is to utilise semantic similarity of publications to identify bridges or brokers in the scholarly communication network. Based on this idea, we developed a method for assessing the

amount of a publication's contribution to the research field and a second method which aims at characterising types of research collaboration to provide an early indication of potential future impacts. We evaluated these methods on several datasets and demonstrated the feasibility of applying these methods in large collections of research publications.

Semantometrics are in the context of this thesis important for a number of reasons. While the idea of utilising publication content for the development of new metrics may seem obvious, text has not received as much attention in research evaluation as other types of data. Despite substantial evidence showing that the existing methods might not be adequate for measuring research publication quality, significant effort is being put into improving the existing research metrics and the existing data instead into developing entirely new approaches. The work presented in this thesis (our TrueImpactDataset) also provides a framework for developing new methods, and we hope our work will inspire further developments in the area of semantometrics. We believe text analysis offers many opportunities for both improving the existing metrics and for developing new metrics, and to demonstrate this point, we have developed two new methods which utilise publication content in a novel way.

Semantometrics, as defined in this thesis, have already received recognition both by the research community and by HEFCE (Higher Education Funding Council for England), who manages the UK REF (Research Excellence Framework). Semantometrics were referenced in a recent book "Research 2.0 and the future of information literacy" [Koltay et al., 2016] and by HEFCE stating that "While many conventional bibliometric approaches are of only limited value [Wilsdon et al., 2015], this new technology [author's note: this is referring to semantometrics] offers the potential to develop truly meaningful measures of research progress."

[Hill, 2016]. Furthermore, the UK not-for-profit organisation Jisc[1], whose role is to support higher education in the UK, recognised the potential of semantometrics by funding two full-time PhD students to continue research on semantometrics.

8.2 Contributions of this thesis

In the previous section we have reiterated our central research question and sub-questions which we tackled in this thesis. Here we summarise how we approached each question and discuss the answers and contributions we brought.

In Chapter 1, Section 1.3, we identified the central research question and the goals of the thesis. The central research question asked how can content be effectively incorporated into evaluation of research publications to provide additional evidence of publication quality and value. We have then broken this question down into sub-questions, which we dealt with in the individual chapters of the thesis. We also set ourselves two goals that further motivated the overall effort and outcomes of the thesis. These goals were focused on (1) designing new methods for assessing the value of research publications and evaluating these methods in comparison with existing research evaluation metrics (Chapters 5, 6 and 7) and (2) showing that the developed metrics can be deployed in large document collections to improve the analysis of published research (Chapters 5 and 7). In this section we provide a summary of the contributions of this thesis to the central research question. The detailed summaries to the separate sub-questions and the research goals can be found in the Conclusions sections at the end of Chapters 3-7.

After providing the background for understanding the research pub-

[1] `https://www.jisc.ac.uk/`

lication evaluation task (Sections 2.1), our first research step explored the existing methods used in evaluation of research publications (Section 2.2). We have broadly categorised the existing methods according to their input data as citation-based (mostly bibliometric methods, Sections 2.2.1 and 2.2.2) and web-based (webometric and altmetric methods, Section 2.2.3). Our review has shown some common limitations of these methods, which stem from the fact both the citation- and the web-based approaches rely on external evidence, particularly the number of interactions in the scholarly communication network. We have then separately focused on a third category (Section 2.2.4) – methods that utilise publication content (including words and keywords as well as full text). We have shown that while a number of researchers have successfully made use of text for various related tasks, significantly fewer studies have focused on developing new methods which utilise text to provide more robust and reliable metrics. Moreover, the existing studies applicable in this area have been largely limited to studying and classifying citation context. Together with the limitations of the citation- and web-based methods, this lack of existing text-based methods constitutes the motivation behind the research work presented in this thesis. The main contribution of Chapter 2 is summarising the existing work in the area of text-based research analysis and evaluation. To the best of our knowledge our review is the first to focus specifically on text-based methods used in research evaluation.

8.2.1 The concept of research publication quality

Considering that our goal was to develop new methods for assessing research publication quality, the first question that we focused on was studying **what is research publication quality** and what factors influence it. In fact, there was a need for us to discover the dimensions of

research publication quality not only to guide our further research, but also to understand what the relation of the existing research evaluation metrics to quality.

The methodology we chose to study this question was composed of two steps. We first performed an in-depth literature review in which we focused on a number of research evaluation frameworks and systems as well as on prior literature on this topic. Namely, we investigated how publication quality is evaluated in five national research evaluation exercises (including in the UK REF and in Australia's ERA) and how it is evaluated in journal peer review. We have also reviewed two previous works which investigated the concept by surveying researchers in the fields of psychology and medicine. In the second step of our investigation, we used an online survey to study the opinion of researchers in different disciplines on which factors contribute the most to research publication quality. The results of the literature review and the survey are reported in Chapter 3.

Our work is among the first to study the concept of research publication quality as perceived by researchers and research evaluation frameworks. While previous works have explored the perspective of journal editors and researchers in specific disciplines, our work is the first to connect and compare the existing studies with national research evaluation exercises performed around the world. This study has revealed that research publications are typically evaluated in terms of three broad criteria: (1) originality (the original contribution the publication provided), (2) rigour (how well was the research performed and the publication written), and (3) significance (what/who did the publication affect). The respondents of our survey viewed particularly rigour as strongly related to publication quality. The reason for this might be that rigour may be easier to judge than originality. This is because deep understanding of the field

may not be necessary to be able to judge a publication according to its rigour. On the other hand, some prior knowledge may be needed to be able to judge originality as well as significance. As the evaluator is rarely as experienced in the field as the author of the research work, evaluating originality is not an easy task to conduct.

The results of this study influenced how we think about publication quality in the later parts of the thesis. In particular, this knowledge has been used in the development of the semantometric contribution measure (Chapter 6) and in the creation of a new dataset for studying research evaluation methods (Chapter 4).

8.2.2 Evaluating research metrics

The next question we targeted was how can we assess the performance of metrics to understand whether they work well. To be able to evaluate the performance of an indicator or a metric, two things are typically needed: (1) a sample of research publications to test the metric on, and (2) a ground truth or reference data to compare the metric with in order to obtain a performance measurement. This question therefore required us to investigate the existing datasets of research publications and methods which are typically used to assess the performance of research metrics. The results of this investigation are reported in Chapter 4. In our investigation of existing publication datasets, we have focused on datasets which are openly available to the research community. A number of recent reports, including "The Metric Tide" report [Wilsdon et al., 2015], have listed openness, transparency, and reproducibility as one of the recommendations for future developments in research evaluation (Chapter 2). Motivated by these recommendations, we have reviewed a number of open datasets of scholarly publications and identified the Microsoft Academic Graph (MAG) as a promising new resource. To inform fu-

ture potential users of the benefits and limitations of the MAG, we have provided both an in-depth analysis of the MAG dataset and a comparison of the MAG against several other external datasets.

Next, we have focused on methods typically used for assessing the performance of research metrics. Our review of this topic revealed a significant issue which, in our opinion, complicates the development of new research evaluation methods. This issue is the lack of evaluation data, such as a ground truth dataset. With this regard, one of the main contributions of this thesis in relation to research evaluation methods in general is that we identified a new approach for analysing the performance of research metrics. Following up on our findings from Chapter 3, we have focused on analysing the performance of research metrics for assessing research contribution. To this end, we have created a dataset consisting of two types of publications – seminal research publications and literature reviews. We have picked these two types of publications as they represent work providing very different amounts of research contribution. The underlying idea behind this dataset is that in research evaluations focused on recognising publications that provided a significant research contribution to their field, seminal papers should on average perform better than literature reviews. This dataset will enable evaluations and analyses of new research metrics, particularly as in this area no ground truth dataset exists.

8.2.3 Beyond citation counting

Once we had an understanding of which factors affect research publication quality and how we can evaluate the performance of research metrics, we used this knowledge to analyse the existing research evaluation metrics. For this analysis we have picked Mendeley reader counts as a representative of altmetrics, citation counts as a representative of bibli-

ometrics, and a number of methods based on citation counts, including the h-index, journal impact, and other metrics. Our goal was to study how well these metrics perform in assessing research publication quality. We have approached this question in two steps and reported our results in Chapter 5.

First, we have evaluated citation counts and Mendeley reader counts using the dataset we developed in answering the second research question (Chapter 4). Our work is the first to provide an evaluation of performance of citation counts and Mendeley reader counts for distinguishing important seminal works from literature reviews. This evaluation has shown that while citation counts distinguish between these two types of publications with a degree of accuracy (63%, i.e. 10% over a random baseline), Mendeley reader counts do not work better than a random baseline on this task for our dataset (highest accuracy 51.05%, while our baseline model achieved 52.87%). We believe this is an important finding which contributes to the discussion on whether citation counts can be used as a proxy to scientific quality [Bornmann and Haunschild, 2017].

In the second step of our investigation of this research question, we have focused on the citation-based metrics. This part of our investigation was conducted through participation in the 2016 WSDM Cup Challenge, in which the submitted publication ranking methods were evaluated against human judgement data [Wade et al., 2016]. The participation in the challenge has therefore enabled us to evaluate the performance of citation counts (including normalised citation counts, the h-index, and other related metrics) against data, which is otherwise difficult to obtain. The goal of the challenge was to assess the importance of research publications using data from the Microsoft Academic Graph (MAG, Chapter 4) and to provide a static rank for publications in the dataset. During this challenge, we focused specifically on various bibliometric methods

and tested over 270 different submissions. In our experiments, we have made several interesting observations about the performance of different metrics, such as the h-index and journal impact, for evaluating individual publications. However, our main contribution to this topic is the demonstration that by combining the information from different types of entities (publications, authors, venues, and affiliations), we can achieve significantly better performance (even without utilising additional data such as altmetrics or text) than by utilising information from a single type of entity at a time. We believe this is an important finding, as it demonstrates simple improvements can be made to the existing research metrics to make these metrics more reliable.

8.2.4 Utilising content for research publication evaluation

Motivated by the possibility of creating more meaningful research evaluation methods which better reflect research publications' quality and by the opportunities provided by the Open Access initiative, we realised that publication content offers an enormous potential for developing new metrics. Therefore, as the final step of our research work, we focused on how to utilise publication content to develop new research evaluation methods that provide more meaningful information related to research publication quality. This work is reported in Chapters 6 and 7. To this end, **we have proposed *semantometrics*, a new class of research evaluation methods which utilise publication content**. In contrast to the existing research metrics which rely on external evidence, semantometrics build on the premise that text is needed to assess the quality of a publication.

To demonstrate the possibilities of semantometrics, we have designed

two new content-based methods for analysing research contribution, which are based on the idea of utilising semantic similarity of publications to identify bridges in the scholarly communication network (Chapter 6). The development of these methods was guided by our findings made in Chapter 3 and was realised utilising datasets presented in Chapter 4. The first method aims at assessing the amount of research contribution a publication provided, and the second method aims at categorising types of research collaboration to provide an early indication of possible future impacts of the publication. While, as we later found out, our *contribution* metric is based on similar underlying assumptions to the method presented by Gerrish and Blei [2010], the specific method, implementation, domain, and our evaluation are new.

We have analysed and evaluated our two semantometric methods on a number of datasets and in comparison to a number of other metrics (Chapters 6 and 7). To analyse our contribution measure and demonstrate it can be deployed in large document collections, we have conducted a comparative evaluation of the measure, in which we compared it with citation counts and Mendeley reader counts (Chapter 7). This evaluation was conducted on a large collection of research publications created by merging three datasets. This evaluation has revealed some interesting and useful properties of the contribution measure. In particular, we have shown that contribution increases with the increasing number of citations; however, after a certain threshold (i.e. for highly cited papers), higher citation counts do not lead on average to a higher contribution. One explanation for this is that receiving more than a certain number of citations reflects the size of the target audience (i.e. visibility of the publication) rather than higher contribution of the underlying research work.

To study whether the specific implementation of our contribution

measure can be improved, we analyse our collaboration categorisation method and provide a comparison of both methods with existing research evaluation metrics; this was done by utilising the dataset that we developed in answering the second research question (Chapter 4). In this evaluation we have studied (alongside of our two semantometric methods) 60 different features describing semantic similarity of publications connected in a scholarly communication network. We have shown that cosine similarity measure [Manning et al., 2008] is a promising function to describe relations between publications in citation networks and between authors in collaboration networks, which helps in distinguishing important seminal publications from literature reviews (Chapter 7).

One of the most important contributions of this thesis is that we were able to show that content based features work better in distinguishing these two types of papers than citation- and web-based measures. To do this we ranked all features according to their accuracy in classifying publications as seminal publications and literature reviews using several different models. Content based features, particularly features describing the breadth of topics contained within each publication's references, ranked high across all models. This is consistent with our intuition that literature reviews tend to reference publications from a wider area than seminal publications, and it also confirms that features describing the breadth of topics contained within a publication's references provide useful information for our contribution measure. More importantly, we were able to confirm that in our dataset, semantic distance between citing and cited publications is higher for seminal publications than for literature reviews. This confirms the underlying assumption our semantometric contribution measure is built on and demonstrates that studying citation patterns in terms of content similarity might provide meaningful information which was not previously available.

8.3 Limitations and future work

In this section, we will discuss the major limitations of our work. We divide these limitations following the narrative of our thesis and for each of them present and discuss ideas for future work which could be used to tackle these limitations.

8.3.1 Publication quality vs. research quality

As we have explained in Chapters 1 and 2, the focus of this thesis was specifically on research publications. We have investigated the concept of research publication quality and used our findings to design new metrics for assessing specific aspects of publication quality. Although research publications arguably represent the main output of research, this may not be the case for all disciplines, and there are other outputs as well as inputs and immediate steps in the research process which deserve the attention of evaluators. Furthermore, the quality of research may not necessarily be captured well in the publications that the research produced. For example, this may happen in case of research which resulted in a patent. A patent might lead to specific societal and economical benefits; however, this may not be visible through research publications associated with that research.

To this end, there are a number of steps which could be taken to extend our work. First, while our investigation of the concept of research publication quality (Chapter 3) has focused on generalisations related to this concept which can be made across all disciplines and publication types, we believe a valuable extension of this work would be providing a comparison of the perception of research publication quality in different disciplines. A similar investigation could also look at the differences between different publication types.

An interesting future direction is applying our semantometric methods to other types of textual research outputs such as patents and books. Because books, in contrast to research publications, tend to be much longer and cover a wider range of topics, one possibility for applying our methods, especially our contribution measure, would be to analyse each chapter separately. While books are typically not cited with a reference to a specific chapter, a semantometric comparison of each chapter with the state-of-the-art combined with an overall evaluation of the book using our contribution measure could provide additional insights into the specific contributions of the book towards the different topics. Furthermore, our method for analysing research collaboration could be particularly useful in the case of patents, where it could be used to categorise inventions by inter-disciplinarity and collaboration emergence and thus facilitate better understanding of the future potential of inventions.

8.3.2 Evaluating research metrics

In Chapter 4 we have discussed methods which are typically used for analysing the performance of research metrics. Most commonly, the analysis is performed either by manual, qualitative examination of results or by comparison with results obtained from another research evaluation metric or metrics. We have seen that in research evaluation, no ground truth dataset which could be used to evaluate new research metrics exists. We see this as a significant issue which complicates further research in this area. To this end we have developed a new dataset of research publications of two types which can be used to analyse the performance of research metrics in their ability to distinguish research publications providing a very different amount of research contribution. While we believe this is an important first step towards developing a reliable method for evaluating the performance of research metrics and a true ground

truth evaluation set, our dataset is still limited in that it focuses on a specific dimension of research publication quality (namely research contribution). As we have explained in the previous section, not only is research a complex process with many inputs and outputs, but also the perception of what contributes to quality may change across disciplines and across different outputs. In Chapter 4 we have discussed requirements which we believe an "ideal" ground truth dataset for evaluating research metrics should satisfy, which we reiterate here. In our view, these requirements are as follows:

- **Multi-disciplinarity:** A dataset containing publications from different scientific areas is important for two reasons. Firstly, publication patterns are different for each discipline, both in terms of productivity and types of outcomes (conference papers, journal papers, books, etc.). This is also important to enable detecting research which finds use outside of its domain.

- **Time span:** The dataset should also contain publications spanning a wider time frame. This is important because publication patterns may change in time.

- **Publication types:** Different types of research publications (e.g. pure research, applied research, literature review, dataset description, etc.) provide different types of impact and should therefore be represented in the dataset.

- **Peer review judgements:** Finally, to provide a reference rank for comparing the research metrics to, the dataset should contain fair and unbiased judgements provided by domain experts. These judgements should rate the publications based on an agreed set of rules and standards.

Creating such a dataset would require significant time and resources, both in terms of collecting a representative sample of publications and in terms of providing peer review judgements for these publications. Providing the peer review judgements could be a common effort and an existing open peer review system could be used for this task. This would require selecting the reference publications, creating a set of rules according to which the papers in the set should be judged, and ensuring fairness of the peer review.

8.3.3 The meaning of a citation

As we have explained in Chapter 2, in bibliometrics and related areas, the use of citations for impact analysis is usually based on the assumption that all citations are equal (have an equal value). Under this assumption a citation from publication a to publication b is interpreted as influence of publication b on publication a. However, it has been shown that acknowledging the influence of prior work is only one of many reasons for citing a publication [Nicolaisen, 2007, Bornmann and Daniel, 2008]. Our work, particularly our contribution measure presented in Chapter 6, alleviates this issue by replacing citations with the semantic distance between the publications citing a paper and the publications cited by a paper. The semantic distance in this case represents how far a field was moved forward thanks to the paper. Our method therefore does not use citations directly for contribution calculation, but rather uses them to identify publications for which to calculate semantic distance.

However, it could be argued that for the identification of publications for calculating semantic distance, our method weights all citations equally. Therefore, a possible future work that we foresee is to combine our approach with the existing citation classification methods. We have reviewed a number of these methods in Chapter 2. We have shown

that the steps involved in these studies include defining a classification scheme, extracting the context of citations from publications, extracting appropriate features from the citation contexts and other parts of the publications, and training a classification model using a dataset of labelled examples (ground truth). There are a number of challenges associated with each of these steps, from identifying implicit citations to collecting labelled examples. Due to these many challenges, this work represents an open research problem in itself which has not yet produced a solution which works well and is applicable in practice. Nevertheless, there are a number of options how this work could be exploited to improve the performance of our methods.

For example, our contribution metrics could be calculated using only "important" citations. Another simpler possibility would be to use a similar approach to the work presented by Bertin et al. [2013], who have studied the distribution of citations found in scientific articles. Only citations found in certain sections of the citing articles (such as in the discussion section) could be used in our contribution metric. An advantage of this approach is that, unlike the citation classification methods, it does not require a labelled set to train a model. As future work we plan to investigate whether calculating our contribution metric utilising only citations found in specific sections improves classification accuracy on our TrueImpactDataset.

8.3.4 Availability of content

Effective use of semantometrics requires unrestricted access to publication content for text and data mining (TDM) purposes. However, we have shown that the access to publication manuscripts is, despite the recent growth of Open Access (OA) publishing, still a significant challenge (Chapter 6). This is the case especially for our contribution metric,

which needs access to publications cited by a given paper and publications citing the paper. Even for OA publications, a significant proportion of references and citing articles may not be openly available, particularly if these references and citing articles are more than a few years old or were published in countries which do not yet support OA publishing.

As a consequence, in our studies presented in Chapter 7, we had to limit ourselves to the use of abstracts. At the moment, there does not seem to be any easy solution to this problem than to rely on full text research publication aggregation systems (for Open Access content) and on the largely limited publisher TDM APIs (for paid access to content). In this regard, as future work it would be valuable to investigate the differences between publication full text, abstracts, and titles for use in different tasks, especially for calculating semantic similarity for our methods. Prior work in this area has investigated the difference between abstracts and the full text of articles and found that a significant proportion of abstracts have at least one sentence in common with the full text [Atanassova et al., 2016]. This is an encouraging result which suggests that abstracts may be used as a suitable replacement of full text where full text is not available.

8.3.5 Extending the contribution metric to evaluate article sets

A useful characteristic of a research article evaluation metric is the ability to extend it to estimate the impact of a group of papers that have something in common. In this section we discuss one possibility how our contribution metric could be extended to article sets, which we would like to investigate further as part of our future work.

Broadly, there are four levels of granularity at which one typically

wants to evaluate impact. We describe these four leves in more detail in Chapter 2. These four levels are (1) individual publications, (2) groups of publications, (3) individual researchers, and (4) groups of researchers. Two common approaches to extending article-level metrics to groups of papers and to researchers are (1) using only the most cited publications and (2) using averaged or weighted citation counts. A typical example of the first method is the h-index [Hirsch, 2005], while an example of the second approach is the Journal Impact Factor [Garfield, 1972]. Other examples of the first approach include the *g-index* [Egghe, 2006] and the *i10-index* (as offered by Google Scholar), and other examples of the second approach include the Eigenfactor [Bergstrom, 2007] and the Scimago Journal Rank [González-Pereira et al., 2010]. These examples are based on the principle of passing article citation counts as the input to a function that produces one value characterising the extended metric. Based on this observation, one possibility for extending article level metrics to estimate the impact of article sets is as follows:

$$contrib\_index = argmax_{P' \subset P}(log1p|P'| \cdot \frac{1}{|P'|} \sum_{p_i \in P'} contrib(p_i)) \quad (8.1)$$

where P denotes the set of articles under evaluation, and the function $log1p$ calculates the natural logarithm of one plus the input value. While the *contrib* function refers to the semantometric contribution function we introduced in Chapter 6, the same principle can also be applied to bibliometric and webometric measures.

The formula combines both approaches to extending bibliometric measures listed above: (1) using averaged citation counts and (2) using most cited publications of the author/venue. The underlying idea is to encourage researchers to focus on *quality rather than quantity.* The formula consists of two parts. The $log1p|'P| \cdot \frac{1}{|P'|} \sum_{p_i \in P'} contrib(p_i)$ part

of the equation represents the average contribution per publication multiplied by the logarithm of the number of publications. The logarithm in the equation causes the index to grow more rapidly with the first few publications, while the growth gradually slows down as the number of publications increases. The second part of the equation is the $argmax_{P' \subset P}$, which means we are looking for a subset of the set of publications that maximise the value. We could say that the metric expresses the average contribution of the best (in terms of contribution value) publications of the set on which it was calculated. The formula therefore encourages quality rather than quantity, which we believe is an important criterion especially when it comes to evaluation of researchers.

Investigating this and other possibilities for extending our methods to evaluate the impacts of article sets is one possible future direction interesting to us.

8.4 Closing remarks

We have opened this thesis with a quote by Vannevar Bush who, among other achievements, organised the Manhattan Project, conceived the National Science Foundation, and in his essay *As We May Think* envisioned a device which inspired the creation of the World Wide Web. Bush helped to convince the American people that the government must support science. Nowadays, research evaluation is a topic which is becoming more and more critical to scientific progress. We expect that the field of research evaluation will continue to grow and will see many new methods being developed. It is our hope that the work presented in this thesis will inspire and facilitate the development of new research evaluation methods which will better reflect research quality than the existing methods, and thus will support science.

Bibliography

Giovanni Abramo, Ciriaco Andrea D'Angelo, and Flavia Di Costa. Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? *Scientometrics*, 84(3):821–833, 2010.

Helmut A Abt and Eugene Garfield. Is the relationship between numbers of references and paper lengths the same for all sciences? *Journal of the American Society for Information Science and Technology*, 53(13): 1106–1112, 2002.

Amjad Abu-Jbara and Dragomir Radev. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 500–509. Association for Computational Linguistics, 2011.

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R Radev. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Hlt-Naacl*, pages 596–606, 2013.

Daniel E Acuna, Stefano Allesina, and Konrad P Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

Jonathan Adams. Collaborations: The rise of research networks. *Nature*, 490(7420):335–336, 2012.

Nancy J Adler and Anne-Wil Harzing. When knowledge wins: Transcending the sense and nonsense of academic rankings. *Academy of Management Learning & Education*, 8(1):72–95, 2009.

Robert Adler, John Ewing, and Peter Taylor. Citation statistics. *Statistical Science*, 24(1):1, 2009.

Shashank Agarwal, Lisha Choubey, and Hong Yu. Automatically classifying the role of citations in biomedical articles. In *AMIA Annual Symposium Proceedings*, volume 2010, page 11. American Medical Informatics Association, 2010.

Isidro F Aguillo, Jose Luis Ortega, and Mario Fernández. Webometric ranking of world universities: Introduction, methodology, and future developments. *Higher education in Europe*, 33(2-3):233–244, 2008.

Dag W Aksnes. Characteristics of highly cited papers. *Research Evaluation*, 12(3):159–170, 2003. doi: 10.3152/147154403781776645.

Dag W Aksnes and Randi Elisabeth Taxt. Peer reviews and bibliometric indicators: a comparative study at a norwegian university. *Research Evaluation*, 13(1):33–41, 2004.

Tomas C. Almind and Peter Ingwersen. Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4):404–426, 1997.

Benjamin M Althouse, Jevin D West, Carl T Bergstrom, and Theodore Bergstrom. Differences in impact factor across fields and over time. *Journal of the Association for Information Science and Technology*, 60 (1):27–34, 2009.

American Association for the Advancement of Science. Historical trends in federal r&d. https://www.aaas.org/page/historical-trends-federal-rd, 2017. Accessed: 2017-05-15.

Mayur Amin and Michael Mabe. Impact factors: use and abuse. *International Journal of Environmental Science and Technology:(IJEST)*, 1(1):1, 2004.

Alessio Ancaiani, Alberto F Anfossi, Anna Barbara, Sergio Benedetto, Brigida Blasi, Valentina Carletti, Tindaro Cicero, Alberto Ciolfi, Filippo Costa, Giovanna Colizza, et al. Evaluating scientific research in italy: The 2004–10 research evaluation exercise. *Research Evaluation*, 24(3):242–255, 2015.

Jens Peter Andersen. *Conceptualising research quality in medicine for evaluative bibliometrics*. PhD thesis, Videnbasen for Aalborg UniversitetVBN, Aalborg UniversitetAalborg University, Aalborg UniversitetsbibliotekAalborg University Library, 2013.

Kristin Antelman. Do open-access articles have a greater research impact? *College & research libraries*, 65(5):372–382, 2004.

John Antonakis, Nicolas Bastardoz, Yonghong Liu, and Chester A Schriesheim. What makes articles highly cited? *The Leadership Quarterly*, 25(1):152–179, 2014.

Salah G Aoun, Bernard R Bendok, Rudy J Rahme, Ralph G Dacey, and H Hunt Batjer. Standardizing the evaluation of scientific and academic performance in neurosurgery—critical review of the "h" index and its variants. *World neurosurgery*, 80(5):e85–e90, 2013.

J Scott Armstrong. Peer review for journals: Evidence on quality control,

fairness, and innovation. *Science and engineering ethics*, 3(1):63–84, 1997.

Douglas N Arnold and Kristine K Fowler. Nefarious numbers. *Notices of the AMS*, 58(3):434–437, 2011.

ArXiv. Arxiv bulk data access. `https://arxiv.org/help/bulk_data`, 2017a. Accessed: 2017-08-12.

ArXiv. Arxiv submission rate statistics. `https://arxiv.org/help/stats/2016_by_area/index`, 2017b. Accessed: 2017-08-12.

Iana Atanassova and Marc Bertin. Temporal properties of recurring in-text references. *D-Lib Magazine*, 22(9/10), 2016.

Iana Atanassova, Marc Bertin, and Vincent Larivière. On the composition of scientific abstracts. *Journal of Documentation*, 72(4):636–647, 2016.

Awais Athar. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session*, pages 81–87. Association for Computational Linguistics, 2011.

Awais Athar and Simone Teufel. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 597–601. Association for Computational Linguistics, 2012a.

Awais Athar and Simone Teufel. Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 18–26. Association for Computational Linguistics, 2012b.

Australian Research Council. Excellence in research for australia: Era 2015 evaluation handbook. Technical report, Australian Government, Australian Research Council, 2015a.

Australian Research Council. State of australian university research: Volume 1 era national report. Technical report, Australian Government, Australian Research Council, 2015b.

Australian Research Council. Excellence in research for australia. `http://www.arc.gov.au/excellence-research-australia`, 2017. Accessed: 2017-07-02.

Philip Ball. Achievement index climbs the ranks, 2007.

Judit Bar-Ilan. Which h-index?—a comparison of wos, scopus and google scholar. *Scientometrics*, 74(2):257–271, 2008.

Mark Bauerlein, Mohamed Gad-el Hak, Wayne Grody, Bill McKelvey, and Stanley W Trimble. We must stop the avalanche of low-quality research. *The Chronicle of Higher Education*, 13, 2010.

C Glenn Begley and John PA Ioannidis. Reproducibility in science. *Circulation research*, 116(1):116–126, 2015.

Carl Bergstrom. Eigenfactor: Measuring the value and prestige of scholarly journals. *C&RL News*, 68(5):314–316, 2007.

Marc Bertin and Iana Atanassova. A study of lexical distribution in citation contexts through the imrad standard. *PloS Negl. Trop. Dis*, 1 (200,920):83–402, 2014.

Marc Bertin and Iana Atanassova. Weak links and strong meaning: The complex phenomenon of negational citations. In *BIR@ ECIR*, pages 14–25, 2016.

Marc Bertin, Iana Atanassova, Vincent Lariviere, and Yves Gingras. The distribution of references in scientific papers: An analysis of the imrad structure. In *Proceedings of the 14th ISSI Conference*, volume 591, page 603, 2013.

Marc Bertin, Iana Atanassova, Yves Gingras, and Vincent Larivière. The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, 67(1):164–177, 2016a.

Marc Bertin, Iana Atanassova, Cassidy R Sugimoto, and Vincent Lariviere. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, 109(3):1417–1434, 2016b.

Sujit Bhattacharya, Hildrun Kretschmer, and Martin Meyer. Characterizing intellectual spaces between science and technology. *Scientometrics*, 58(2):369–390, 2003.

Jin Bihui, Liang Liming, Ronald Rousseau, and Leo Egghe. The r-and ar-indices: Complementing the h-index. *Chinese science bulletin*, 52 (6):855–863, 2007.

D a Bini, G M Del Corso, and F Romani. Evaluating Scientific Products by Means of Citation-Based Models: A First Analysis and Validation. *Electronic Transactions on Numerical Analysis*, 33:1–16, 2008.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, Yee Fan Tan, et al. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*, 2008.

Maria Biryukov and Cailing Dong. Analysis of computer science communities based on dblp. *Research and advanced technology for digital libraries*, pages 228–235, 2010.

Bo-Christer Björk, Patrik Welling, Mikael Laakso, Peter Majlender, Turid Hedlund, and Gudni Gudnason. Open access to the scientific journal literature: situation 2009. *PloS one*, 5(6):e11273, January 2010.

Samuel Bjork, Avner Offer, and Gabriel Söderberg. Time series citation data: the nobel prize in economics. *Scientometrics*, 98(1):185–196, 2014.

Lennart Björneborn and Peter Ingwersen. Toward a basic framework for webometrics. *Journal of the American society for information science and technology*, 55(14):1216–1227, 2004.

Johan Bollen, Herbert Van de Sompel, Aric Hagberg, and Ryan Chute. A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE*, 4(6):e6022, 2009.

Katy Börner, Chaomei Chen, and Kevin W Boyack. Visualizing knowledge domains. *Annual review of information science and technology*, 37(1):179–255, 2003.

Lutz Bornmann. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of informetrics*, 8(4):895–903, 2014.

Lutz Bornmann. Usefulness of altmetrics for measuring the broader impact of research. *Aslib Journal of Information Management*, 67(3):305, 2015.

Lutz Bornmann and Hans-Dieter Daniel. Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3):391–392, 2005.

Lutz Bornmann and Hans-Dieter Daniel. Selecting scientific excellence through committee peer review-a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, 68(3):427–440, 2006.

Lutz Bornmann and Hans-Dieter Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008.

Lutz Bornmann and Robin Haunschild. Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? *Scientometrics*, 110(2):937–943, 2017.

Lutz Bornmann and Loet Leydesdorff. Does quality and content matter for citedness? a comparison with para-textual factors and over time. *Journal of Informetrics*, 9(3):419–429, 2015.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.

Lutz Bornmann, Irina Nast, and Hans-Dieter Daniel. Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? a quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication. *Scientometrics*, 77(3):415–432, 2008.

Robert R Braam, Henk F Moed, and Anthony FJ Van Raan. Mapping of science by combined co-citation and word analysis i. structural aspects. *Journal of the American Society for information science*, 42(4):233, 1991.

Samuel C. Bradford. Sources of information on specific subjects. *Engineering*, 137(3550):85–86, 1934.

Maria Bras-Amorós, Josep Domingo-Ferrer, and Vicenç Torra. A bibliometric index based on collaboration distances. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 5–6. Springer, 2010.

Tibor Braun, Wolfgang Glänzel, and András Schubert. A hirsch-type index for journals. *Scientometrics*, 69(1):169–173, 2006.

Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th International World-Wide Web Conference*, page 20, 1998.

Tim Brody, Stevan Harnad, and Leslie Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the Association for Information Science and Technology*, 57(8):1060–1072, 2006.

Patrick O. Brown, Diane Cabell, Aravinda Chakravarti, Barbara Cohen, Tony Delamothe, Michael Eisen, Les Grivell, Jean-Claude Guedon, R. Scott Hawley, Richard K. Johnson, Marc W. Kirschner, David Lipman, Arnold P. Lutzker, Elizabeth Marincola, Richard J. Roberts, Gerald M. Rubin, Robert Schloegl, Vivian Siegel, Anthony D. So, Peter Suber, Harold E. Varmus, Jan Velterop, Mark J. Walport, and Linda Watson. Bethesda Statement on Open Access Publishing. `http://legacy.earlham.edu/~peters/fos/bethesda.htm`, 6 2003. Accessed: 2017-09-11.

Roger A Brumback. Impact factor wars: Episode v—the empire strikes back, 2009.

Hans-Jorg Bullinger, Karl Max Einhaupl, Peter Gaehtgens, Peter Gruss, Hans-Olaf Henkel, Walter Kroll, Ernst-Ludwig Winnacker, Bernard Larrouturou, Jurgen Mittelstrass, Paolo Galluzzi, Christian Brechot, Yehuda Elkana, Jean-Claude Guedon, Martin Roth, Friedrich Geisselmann, Jose Miguel Ruano Leon, Dieter Simon, Jens Braarvig, and Peter Schirmbacher. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. `http://openaccess.mpg.de/286432/Berlin-Declaration`, 10 2003. Accessed: 2017-09-11.

John Burns, Alan Brenner, Keith Kiser, Michael Krot, Clare Llewellyn, and Ronald Snyder. Jstor: Data for research. *European Conference on Research and Advanced Technology for Digital Libraries*, pages 416–419, 2009.

Linda Butler. Using a balanced approach to bibliometrics: quantitative performance measures in the australian research quality framework. *Ethics in Science and Environmental Politics*, 8(1):83–92, 2008.

Bilal Hayat Butt, Muhammad Rafi, Arsal Jamal, Raja Sami Ur Rehman, Syed Muhammad Zubair Alam, and Muhammad Bilal Alam. Classification of research citations (crc). *arXiv preprint arXiv:1506.08966*, 2015.

Michel Callon, Jean-Pierre Courtial, William A Turner, and Serge Bauin. From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, 22 (2):191–235, 1983.

Michel Callon, Jean-Pierre Courtial, and Francoise Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1):155–205, 1991.

Juan Campanario. Rejecting and resisting nobel class discoveries: accounts by nobel laureates. *Scientometrics*, 81(2):549–565, 2009.

Juan Miguel Campanario and Erika Acedo. Rejecting highly cited papers: The views of scientists who encounter resistance to their discoveries from other scientists. *Journal of the American Society for Information Science and Technology*, 58(5):734–743, 2007.

Philip Campbell. Escape from the impact factor. *Ethics in science and environmental politics*, 8(1):5–7, 2008.

Cornelia Caragea, Jian Wu, Alina Maria Ciobanu, Kyle Williams, Juan Pablo Fernández Ramírez, Hung-Hsuan Chen, Zhaohui Wu, and C Lee Giles. Citeseerx: A scholarly big dataset. In *ECIR*, volume 14, pages 311–322. Springer, 2014.

Cornelia Caragea, Florin Adrian Bulgarov, and Rada Mihalcea. Co-training for topic classification of scholarly data. In *EMNLP*, pages 2357–2366, 2015.

Davide Castelvecchi. Physics Paper Sets Record with More than 5,000 Authors. *Nature News*, pages 14–17, 2015. ISSN 1476-4687. doi: 10.1038/nature.2015.17567. URL `http://www.nature.com/doifinder/10.1038/nature.2015.17567`.

Timothy Caulfield, Shawn HE Harmon, and Yann Joly. Open science versus commercialization: a modern research conflict? *Genome medicine*, 4(2):17, 2012.

Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-*

CS Joint Conference on Digital Libraries, pages 351–360. IEEE Press, 2014.

Leslie Chan, Darius Cuplinskas, Michael Eisen, Fred Friend, Yana Genova, Jean-Claude Guedon, Melissa Hagemann, Stevan Harnad, Rick Johnson, Rima Kupryte, Manfredi La Manna, Istvan Rév, Monika Segbert, Sidnei de Souza, Peter Suber, and Jan Velterop. Budapest Open Access Initiative. `http://www.budapestopenaccessinitiative.org/read`, 2 2002. Accessed: 2017-09-11.

Shu-Kai Chang, Sui-Tsung Go, Yueh-Hua Wu, Yen-Ting Lee, Chien-Lin Lai, Sz-Han Yu, Chun-Wei Chen, Huan-Yuan Chen, Ming-Feng Tsai, Mi-Yen Yeh, and Shou-De Lin. An ensemble of ranking strategies for static rank prediction in a large heterogeneous graph. In *Proceedings of WSDM Cup 2016 – Entity Ranking Challenge at the 9th ACM International Conference on Web Search and Data Mining (WSDM 2016)*, 2016.

CiteSeerX. Citeseerx data. `http://csxstatic.ist.psu.edu/about/data`, 2017. Accessed: 2017-08-09.

Clarivate Analytics. InCites Help: Cited Half-life. `http://ipscience-help.thomsonreuters.com/inCites2Live/indicatorsGroup/aboutHandbook/usingCitationIndicatorsWisely/citedHalfLife.html`, 2017a. Accessed: 2017-09-07.

Clarivate Analytics. InCites Help: Immediacy index. `http://ipscience-help.thomsonreuters.com/inCites2Live/indicatorsGroup/aboutHandbook/usingCitationIndicatorsWisely/immediacyIndex.html`, 2017b. Accessed: 2017-09-07.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Manolo J Cobo, Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the Association for Information Science and Technology*, 62(7):1382–1402, 2011.

Francis J. Cole and Nellie B. Eales. The history of comparative anatomy. part i: A statistical analysis of the literature. *Science Progress*, 11(43): 578–596, 1917.

Cristian Colliander. A novel approach to citation normalization: A similarity-based method for creating reference sets. *Journal of the Association for Information Science and Technology*, 66(3):489–500, 2015.

Consejo Superior de Investigaciones Científicas. Ranking Web of Universities: 2015 Edition. `http://www.webometrics.info/en`, 2015. URL `http://www.webometrics.info/en`. Accessed: 2016-04-29.

CORE: Connecting Repositories. Services. `https://core.ac.uk/services`, 2017. Accessed: 2017-08-12.

Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. Do "altmetrics" correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10):2003–2019, 2015.

Blaise Cronin and Gail McKenzie. The trajectory of rejection. *Journal of Documentation*, 48(3):310–317, 1992.

Edit Csajbók, Anna Berhidi, Lívia Vasas, and András Schubert. Hirsch-index for countries based on essential science indicators data. *Scientometrics*, 73(1):91–117, 2007.

Lei Cui. Rating Health Web sites using the principles of Citation Analysis: A Bibliometric Approach. *Journal of Medical Internet Research*, 1(1):e4, 1999.

Ciriaco Andrea D'Angelo and Giovanni Abramo. Publication rates in 192 research fields of the hard sciences. In *Proceedings of the 15th ISSI Conference*, pages 915–925, 2015.

Philip Davis and Michael Fromerth. Does the arxiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2):203–215, 2007.

Philip M Davis. The persistence of error: a study of retracted articles on the internet and in personal libraries. *Journal of the Medical Library Association*, 100(3):184, 2012.

Nicola De Bellis. *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Scarecrow Press, 2009.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Gianna M Del Corso and Francesco Romani. A time-aware citation-based model for evaluating scientific products. In *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, page 15. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

Department of Energy Office of Science. Doe bes budget and operational review of the sns and hfir, 2014. URL http://neutrons.ornl.gov/sites/default/files/DOE%20SNS-HFIR%20Cost%20Review%20November%202013.pdf.

Angelo Di Iorio, Andrea Giovanni Nuzzolese, and Silvio Peroni. Towards the automatic identification of the nature of citations. In *SePublica*, pages 63–74, 2013.

Ivan Díaz, Martí Cortey, Àlex Olvera, and Joaquim Segalés. Use of h-index and other bibliometric indicators to evaluate research productivity outcome on swine diseases. *PloS one*, 11(3):e0149690, 2016.

Ying Ding, Gobinda G Chowdhury, and Schubert Foo. Bibliometric cartography of information retrieval research by using co-word analysis. *Information processing & management*, 37(6):817–842, 2001.

Ying Ding, Xiaozhong Liu, Chun Guo, and Blaise Cronin. The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, 7(3):583–592, 2013.

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9):1820–1833, 2014.

Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. Will this paper increase your h-index?: Scientific impact prediction. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 149–158. ACM, 2015.

Suhendry Effendy and Roland HC Yap. Analysing trends in computer science research: A preliminary study using the microsoft academic

graph. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1245–1250. International World Wide Web Conferences Steering Committee, 2017.

Leo Egghe. An improvement of the h-index: the g-index. *ISSI newsletter*, 2(1):8–9, 2006.

Leo Egghe and Ronald Rousseau. *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers, 1990.

Mojisola Erdt, Aarthy Nagarajan, Sei-Ching Joanna Sin, and Yin-Leng Theng. Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, 109(2):1117–1166, 2016.

Gunther Eysenbach. Citation advantage of open access articles. *PLoS biology*, 4(5):e157, 2006.

Mohammad A Abolghassemi Fakhree and Abolghasem Jouyban. Scientometric analysis of the major iranian medical universities. *Scientometrics*, 87(1):205–220, 2011.

Matthew E Falagas, Eleni I Pitsouni, George A Malietzis, and Georgios Pappas. Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342, 2008.

Jia Feng, Yun Qiu Zhang, and Hao Zhang. Improving the co-word analysis method based on semantic distance. *Scientometrics*, 111(3):1521–1531, 2017.

Ming-Han Feng, Kuan-Hou Chan, Huan-Yuan Chen, Ming-Feng Tsai, Mi-Yen Yeh, and Shou-De Lin. An efficient solution to reinforce paper

ranking using author/venue/citation information – the winner's solution for wsdm cup 2016. In *Proceedings of WSDM Cup 2016 – Entity Ranking Challenge at the 9th ACM International Conference on Web Search and Data Mining (WSDM 2016)*, 2016.

Dalibor Fiala. Mining citation information from citeseer data. *Scientometrics*, 86(3):553–562, 2011.

Dalibor Fiala. Bibliometric analysis of citeseer data for countries. *Information Processing & Management*, 48(2):242–253, 2012.

Massimo Franceschet. Journal influence factors. *Journal of Informetrics*, 4(3):239–248, 2010.

Fiorenzo Franceschini and Domenico Maisano. Critical remarks on the italian research assessment exercise vqr 2011–2014. *Journal of Informetrics*, 11(2):337–357, 2017.

Fiorenzo Franceschini, Domenico Maisano, and Luca Mastrogiacomo. The museum of errors/horrors in scopus. *Journal of Informetrics*, 10 (1):174–182, 2016.

Olivier Francois. Arbitrariness of peer review: A bayesian analysis of the nips experiment. *arXiv preprint arXiv:1507.06411*, 2015.

Finbar Galligan and Sharon Dyas-Correia. Altmetrics: Rethinking the Way We Measure. *Serials Review*, 39(1):56–61, mar 2013.

Eugene Garfield. Citation indexes for science. *Science*, 122:108–111, 1955.

Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

Eugene Garfield. The meaning of the impact factor. *International Journal of Clinical and Health Psychology*, 3(2), 2003.

Eugene Garfield. The History and Meaning of the Journal Impact Factor. *JAMA: the journal of the American Medical Association*, 295(1):90–93, 2006.

Eugene Garfield et al. Science citation index-a new dimension in indexing. *Science*, 144(3619):649–654, 1964.

Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

Sean Gerrish and David M Blei. A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 375–382, 2010.

Aldo Geuna and Ben R Martin. University research evaluation and funding: An international comparison. *Minerva*, 41(4):277–304, 2003.

C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98. ACM, 1998.

Wolfgang Glänzel, Balázs Schlemmer, and Bart Thijs. Better late than never? on the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3):571–586, 2003.

Wolfgang Glänzel, Sarah Heeffer, and Bart Thijs. Lexical analysis of scientific publications for nano-level scientometrics. *Scientometrics*, pages 1–10, 2017.

Patrick Glenisson, Wolfgang Glänzel, and Olle Persson. Combining full-text analysis and bibliometric indicators. a pilot study. *Scientometrics*, 63(1):163–180, 2005.

Borja González-Pereira, Vicente P Guerrero-Bote, and Félix Moya-Anegón. A new approach to the metric of journals' scientific prestige: The sjr indicator. *Journal of informetrics*, 4(3):379–391, 2010.

Roger Guimerà, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(April):697–702, 2005. ISSN 0036-8075. doi: 10.1126/science.1106340.

Susan Guthrie, Watu Wamae, Stephanie Diepeveen, Steven Wooding, and Jonathan Grant. *Measuring research: a guide to research evaluation frameworks and tools*. RAND Europe, 2013.

David Hall, Daniel Jurafsky, and Christopher D Manning. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics, 2008.

David P Hamilton. who's uncited now? *Science*, 251(4989):25–26, 1991.

Stevan Harnad and Tim Brody. Comparing the impact of open access (oa) vs. non-oa articles in the same journals. *D-lib Magazine*, 10(6), 2004.

Gareth Harries, David Wilkinson, Liz Price, Ruth Fairclough, and Mike Thelwall. Hyperlinks as a data source for science mapping. *Journal of Information Science*, 30(5):436–447, October 2004.

Nigel Harwood. An interview-based study of the functions of citations

in academic writing across two disciplines. *Journal of Pragmatics*, 41 (3):497–518, 2008.

Anne-Wil Harzing. Document categories in the isi web of knowledge: Misunderstanding the social sciences? *Scientometrics*, 94(1):23–34, 2013.

Anne-Wil Harzing. Microsoft academic (search): a phoenix arisen from the ashes? *Scientometrics*, 108(3):1637–1647, 2016.

Anne-Wil Harzing and Satu Alakangas. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016.

Saeed-Ul Hassan and Peter Haddawy. Measuring international knowledge flows and scholarly impact of scientific research. *Scientometrics*, 94(1): 163–179, 2013.

Ellen Hazelkorn et al. Assessing europe's university-based research. Technical report, European Commission Directorate-General for Research, 2010.

Qin He. Knowledge discovery through co-word analysis. *Library trends*, 48(1):133, 1999.

Victor Henning and Jan Reichelt. Mendeley-a last. fm for research? In *eScience, 2008. eScience'08. IEEE Fourth International Conference on*, pages 327–328. IEEE, 2008.

Diana Hicks. Performance-based university research funding systems. *Research policy*, 41(2):251–261, 2012.

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. The leiden manifesto for research metrics. *Nature*, 520(7548): 429, 2015.

Steven A Hill. Making the future of scholarly communications. *Learned Publishing*, 29(S1):366–370, 2016.

Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, pages 16569–16572, 2005.

Dirk Holste, Ivana Roche, Marianne Hörlesberger, Dominique Besagni, Thomas Scherngell, Claire Francois, Pascal Cuxac, and Edgar L Schiebel. A concept for inferring'frontier research'in research project proposals. In *13th International Conference on Scientometrics and Informetrics*, 2011.

William W. Hood and Concepción S. Wilson. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291–314, 2001.

Marianne Hörlesberger, Ivana Roche, Dominique Besagni, Thomas Scherngell, Claire François, Pascal Cuxac, Edgar Schiebel, Michel Zitt, and Dirk Holste. A concept for inferring 'frontier research' in grant proposals. *Scientometrics*, 97(2):129–148, 2013.

Wen-Ru Hou, Ming Li, and Deng-Ke Niu. Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays*, 33(10):724–727, 2011.

Chin-Chi Hsu, Kuan-Hou Chan, Ming-Han Feng, Yueh-Hua Wu, Huan-Yuan Chen, Sz-Han Yu, Chun-Wei Chen, Ming-Feng Tsai, Mi-Yen Yeh, and Shou-De Lin. Time-aware weighted pagerank for paper ranking in academic graphs. In *Proceedings of WSDM Cup 2016 – Entity Ranking Challenge at the 9th ACM International Conference on Web Search and Data Mining (WSDM 2016)*, 2016.

Zhigang Hu, Chaomei Chen, and Zeyuan Liu. The recurrence of citations within a scientific article. In *ISSI*, 2015.

Sven E Hug and Martin P Brändle. The coverage of microsoft academic: Analyzing the publication output of a university. *arXiv preprint arXiv:1703.05539*, 2017.

B Ian Hutchins, Xin Yuan, James M Anderson, and George M Santangelo. Relative citation ratio (rcr): A new metric that uses citation rates to measure influence at the article level. *PLoS Biol*, 14(9):e1002541, 2016.

Peter Ingwersen. The Calculation of Web Impact Factors. *Journal of Documentation*, 54(2):236–243, 1998.

John PA Ioannidis. How to make more published research true. *PLoS Med*, 11(10):e1001747, 2014.

Yoo Kyung Jeong, Min Song, and Ying Ding. Content-based author co-citation analysis. *Journal of Informetrics*, 8(1):197–211, 2014.

Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R Radev. Nlp-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1):93–130, 2017.

Arif E Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Citation classification for behavioral analysis of a scientific field. *arXiv preprint arXiv:1609.00435*, 2016.

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and C Lee Giles. Context sensitive topic models for author influence in document net-

works. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, page 2274, 2011.

Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. Peer review in scientific publications: benefits, critiques, & a survival guide. *Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine*, 25(3):227, 2014.

M. M. Kessler. Bibliographic Coupling Between Scientific Papers. *American Documentation*, 14(1):10–25, 1963.

Madian Khabsa and C Lee Giles. The number of scholarly documents on the public web. *PloS one*, 9(5):e93949, 2014.

Hamidreza Kianifar, Ramin Sadeghi, and Leili Zarifmahmoudi. Comparison between impact factor, eigenfactor metrics, and scimago journal rank indicator of pediatric neurology journals. *Acta Informática Médica*, 22(2):103, 2014.

Stefan Klampfl and Roman Kern. An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles. In *International Conference on Theory and Practice of Digital Libraries*, pages 144–155. Springer, 2013.

Alex Knapp. How much does it cost to find a higgs boson? `https://www.forbes.com/sites/alexknapp/2012/07/05/how-much-does-it-cost-to-find-a-higgs-boson/#3948092c3948`, 2012. Accessed: 2017-05-17.

Petr Knoth and Zdenek Zdrahal. Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12), 2012.

Tibor Koltay, Sonja Spiranec, and László Z Karvalics. *Research 2.0 and the future of information literacy*. Chandos Publishing, 2016.

Ronald N Kostoff, J Antonio del Rio, James A Humenik, Esther Ofilia García, and Ana María Ramírez. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the Association for Information Science and Technology*, 52(13):1148–1156, 2001.

Gabriel Kreiman and John HR Maunsell. Nine criteria for a measure of scientific output. *Frontiers in computational neuroscience*, 5(48):11, 2011.

Grzegorz Kreiner. The slavery of the h-index—measuring the unmeasurable. *Frontiers in human neuroscience*, 10, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Renaud Lambiotte and Pietro Panzarasa. Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3):180–190, 2009.

Mark A Largent and Julia I Lane. Star metrics and the science of science policy. *Review of Policy Research*, 29(3):431–438, 2012.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Kai Eckert. Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*. ACM, 2017.

Themis Lazaridis. Ranking university departments using the mean h-index. *Scientometrics*, 82(2):211–216, 2010.

Bangrae Lee and Yong-Il Jeong. Mapping korea's national r&d domain of robot technology by using the co-word analysis. *Scientometrics*, 77 (1):3–19, 2008.

Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013.

Yen-Chun Lee, Chaomei Chen, and Xing-Tzu Tsai. Visualizing the knowledge domain of nanoparticle drug delivery technologies: A scientometric review. *Applied Sciences*, 6(1):11, 2016.

Michael Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *String processing and information retrieval*, pages 481–486. Springer, 2002.

Michael Ley. Dblp xml requests. `http://dblp.uni-trier.de/xml/docu/dblpxmlreq.pdf`, 2009. Accessed: 2017-08-10.

Loet Leydesdorff. Words and co-words as indicators of intellectual organization. *Research Policy*, 18(4):209–223, 1989.

Loet Leydesdorff and Michael Curran. Mapping University-Industry-Government Relations on the Internet: The Construction of Indicators for a Knowledge-Based Economy. *Cybermetrics*, 4(1):1–17, 2003.

Loet Leydesdorff and Iina Hellsten. Metaphors and diaphors in science communication: Mapping the case of stem cell research. *Science communication*, 27(1):64–99, 2005.

Xiang Li, Yifan He, Adam Meyers, and Ralph Grishman. Towards fine-grained citation function classification. In *RANLP*, pages 402–407, 2013a.

Xuemei Li and Mike Thelwall. F1000, mendeley and traditional bibliometric indicators. In *Proceedings of the 17th international conference on science and technology indicators*, volume 2, pages 451–551, 2012.

Yunrong Li, Filippo Radicchi, Claudio Castellano, and Javier Ruiz-Castillo. Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, 7(3):746–755, 2013b.

Shengbo Liu, Kun Ding, Bo Wang, Delong Tang, and Zhao Qu. The research of paper influence based on citation context - a case study of the nobel prize winner's paper. In *ISSI*, 2015.

Xiaozhong Liu, Jinsong Zhang, and Chun Guo. Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the Association for Information Science and Technology*, 64(9):1852–1863, 2013.

Avishay Livne, Eytan Adar, Jaime Teevan, and Susan Dumais. Predicting citation counts using text and graph mining. In *Proc. the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*, 2013.

Alfred J. Lotka. The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, 16:317–323, 1926.

Dongsheng Luo, Chen Gong, Renjun Hu, Liang Duan, and Shuai Ma. Ensemble enabled weighted pagerank. In *Proceedings of WSDM Cup 2016 – Entity Ranking Challenge at the 9th ACM International Conference on Web Search and Data Mining (WSDM 2016)*, 2016.

Michael H MacRoberts and Barbara R MacRoberts. Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, 61(1): 1–12, 2010.

Nabeil Maflahi and Mike Thelwall. When are readership counts as useful as citation counts? scopus versus mendeley for lis journals. *Journal of the Association for Information Science and Technology*, 67(1):191–199, 2016. DOI: 10.1002/asi.23369.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*. Cambridge university press Cambridge, 2008.

Katherine W. McCain. Mapping Economics through the Journal Literature: An Experiment in Journal Cocitation Anaiysis. *Journal of the American Society for Information Science*, 42(4):290–296, 1991.

Kathy McKeown, Hal Daume, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R Fleischmann, et al. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696, 2016.

Marie E McVeigh and Stephen J Mann. The journal impact factor denominator: defining citable (counted) items. *Jama*, 302(10):1107–1109, 2009.

Lokman I Meho. The rise and rise of citation analysis. *Physics World*, 20(1):32, 2007.

Robert K Merton. Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society*, 105(5):470–486, 1961.

Robert K. Merton. The Matthew Effect in Science. *Science*, 159(3810): 56–63, 1968.

Microsoft Azure. Cognitive services pricing – academic knowledge api. `https://azure.microsoft.com/en-us/pricing/details/cognitive-services/academic-knowledge-api/`, 2017. Accessed: 2017-08-13.

Microsoft Research. Wsdm cup challenge. `https://wsdmcupchallenge.azurewebsites.net/`, 2015. Accessed: 2017-08-20.

Microsoft Research. Kdd cup 2016 – whose papers are accepted the most: towards measuring the impact of research institutions. `https://kddcup2016.azurewebsites.net/`, 2016. Accessed: 2017-08-13.

Microsoft Research. Microsoft academic graph download site. `https://academicgraph.blob.core.windows.net/graph/index.html`, 2017. Accessed: 2017-08-13.

Henk F Moed. Measuring contextual citation impact of scientific journals. *Journal of informetrics*, 4(3):265–277, 2010.

Henk F. Moed. The Source Normalized Impact per Paper Is a Valid and Sophisticated Indicator of Journal Citation Impact. *Journal of the American Society for Information Science and Technology*, 62(1): 211–213, 2011.

Sergio Lopez Montolio, David Dominguez-Sal, and Josep Lluis Larriba-Pey. Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, 42(2):11–16, 2013. ISSN 01635808. doi: 10.1145/2503792.2503795.

Stuart Mullins. Measures, metrics, and indicators. `http://it.toolbox.com/blogs/dw-cents/measures-metrics-and-indicators-23543`, 2009. Accessed: 2017-09-29.

Vassily V. Nalimov and Z. M. Mulchenko. Naukometriya. izuchenie razvitiya nauki kak informatsionnogo protsessa. [scientometrics. study of the development of science as an information process]. *Nauka. (English translation: 1971. Washington, DC: Foreign Technology Division. US Air Force Systems Command, Wright-Patterson AFB, Ohio. (NTIS Report No. AD735-634))*, 34:107–247, 1969.

National Agency for the Evaluation of Universities and Research Institutes. Evaluation of research quality 2011–2014 (vqr 2011–2014). `http://www.anvur.org/attachments/article/825/Call%20VQR_2011_2014_11nov_~.pdf`, 2015. Accessed: 2017-07-01.

Nature Neuroscience Editors. Editorial: Pros and cons of open peer review. *Nature Neuroscience*, 2(3), 1999.

Olgica Nedić and Aleksandar Dekanski. Priority criteria in peer review of scientific articles. *Scientometrics*, 107(1):15–26, 2016.

Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101 (suppl 1):5200–5205, 2004.

David Nicholas, Anthony Watkinson, Hamid R Jamali, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine.

Peer review: still king in the digital age. *Learned Publishing*, 28(1): 15–21, 2015.

Jeppe Nicolaisen. Citation analysis. *Annual review of information science and technology*, 41(1):609–641, 2007.

Ed CM Noyons and Anthony FJ van Raan. Bibliometric cartography of scientific and technological developments of an r & d field. *Scientometrics*, 30(1):157–173, 1994.

Eric Oberesch and Sven Groppe. The mf-index: A citation-based multiple factor index to evaluate and compare the output of scientists. *Open Journal of Web Technologies (OJWT)*, 4(1):1–32, 2017. ISSN 2199-188X. URL `https://www.ronpub.com/OJWT_2017v4i1n01_Oberesch.pdf`.

Office of Science and Technology Policy. Increasing access to the results of federally funded scientific research. `https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf`, 2013. Accessed: 2017-09-11.

Natsuo Onodera and Fuyuki Yoshikane. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4):739–764, 2015.

Omwoyo Bosire Onyancha and Dennis N Ocholla. An informetric investigation of the relatedness of opportunistic infections to hiv/aids. *Information Processing & Management*, 41(6):1573–1588, 2005.

Jan Oosterhaven. Too many journals? towards a theory of repeated rejections and ultimate acceptance. *Scientometrics*, 103(1):261–265, 2015.

Enrique Orduña-Malea, Juan M Ayllón, Alberto Martín-Martín, and Emilio Delgado López-Cózar. Methods for estimating the size of google scholar. *Scientometrics*, 104(3):931–949, 2015.

Judith M Panitch and Sarah Michalak. The serials crisis. *A White Paper for the UNC-Chalep Hill Scholarly Communications Convocation. Janury*, 2005.

Jihyun Park, Margaret Blume-Kohout, Ralf Krestel, Eric T Nalisnick, and Padhraic Smyth. Analyzing nih funding patterns over time with statistical text analysis. In *AAAI Workshop: Scholarly Big Data*, 2016.

Oren Patashnik. *Bibtexing*, 1988.

Robert M Patton, Christopher G Stahl, Thomas E Potok, and Jack C Wells. Identification of user facility related publications. *D-Lib Magazine*, 18(7/8), 2012.

Robert M Patton, Christopher G Stahl, and Jack C Wells. Measuring scientific impact beyond citation counts. *D-Lib Magazine*, 22(9/10):5, 2016.

H Peters and A Van Raan. Structuring scientific activities by co-author analysis: An expercise on a university faculty level. *Scientometrics*, 20 (1):235–255, 1991.

HPF Peters and Anthony FJ van Raan. Co-word-based science maps of chemical engineering. part ii: Representations by combined clustering and multidimensional scaling. *Research Policy*, 22(1):47–71, 1993.

Heather Piwowar. Altmetrics: Value all research products. *Nature*, 493 (7431):159–159, 2013.

Heather Piwowar and Jason Priem. The power of altmetrics on a cv. *Bulletin of the American Society for Information Science and Technology*, 39(4):10–13, 2013. DOI: 10.1002/bult.2013.1720390405.

Derek J. de Solla Price. Networks of Scientific Papers. *Science*, 149 (3683):510–515, 1965.

Derek J. de Solla Price. Citation measures of hard science, soft science, technology, and nonscience. *Communication among scientists and engineers*, pages 3–22, 1970.

Derek J. de Solla Price. A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*, 27(5-6):292–306, 1976.

Derek J. de Solla Price. Multiple authorship. *Science*, 212(4498):986–986, 1981.

Derek J. de Solla Price. *Little science, big science... and beyond*. Columbia University Press New York, 1986.

David Pride and Petr Knoth. Incidental or influential? challenges in automatically detecting citation importance using publication full texts. In *International Conference on Theory and Practice of Digital Libraries*, pages 572–578. Springer, 2017.

Jason Priem. Altmetrics. In Blaise Cronin and Cassidy R Sugimoto, editors, *Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact*, chapter 14, pages 263–288. MIT Press, Cambridge, MA, 2014.

Jason Priem and Bradely H Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday*, 15(7), 2010.

Jason Priem, Dario Taraborelli, Paul Groth, and Cameron Neylon. Altmetrics: A manifesto. 2010.

Alan Pritchard. Statistical bibliography or bibliometrics. *Journal of documentation*, 25:348, 1969.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944, 2013.

Dragomir R Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. A bibliometric and network analysis of the field of computational linguistics. *Journal of the American Society for Information Science and Technology*, 67:683–706, 2016.

Radoslav Radoulov. Exploring automatic citation classification. Master's thesis, University of Waterloo, 2008.

Florian Reitz and Oliver Hoffmann. An analysis of the evolving coverage of computer science sub-fields in the dblp digital library. In *International Conference on Theory and Practice of Digital Libraries*, pages 216–227. Springer, 2010.

Research Councils UK. Rcuk policy on open access and supporting guidance. `http://www.rcuk.ac.uk/documents/documents/rcukopenaccesspolicy-pdf/`, 2012. Accessed: 2017-09-11.

Research Excellence Framework. Panel criteria and working methods. Technical report, Higher Education Funding Council for England, 2012.

Research Excellence Framework. Research Excellence Framework (REF) 2014 Units of Assessment. `http://www.ref.ac.uk/panels/unitsofassessment/`, 12 2014a. Accessed: 2016-11-11.

Research Excellence Framework. Research excellence framework 2014: The results. Technical report, Higher Education Funding Council for England, 2014b.

Diana Rhoten and Walter W Powell. The frontiers of intellectual property: Expanded protection versus new models of open science. *Annu. Rev. Law Soc. Sci.*, 3:345–373, 2007.

Sabir Ribas, Alberto Ueda, Rodrygo LT Santos, Berthier Ribeiro-Neto, and Nivio Ziviani. Simplified relative citation ratio for static paper ranking: Ufmg/latin at wsdm cup 2016. In *Proceedings of WSDM Cup 2016 – Entity Ranking Challenge at the 9th ACM International Conference on Web Search and Data Mining (WSDM 2016)*, 2016.

Martin Ricker. Letter to the editor: About the quality and impact of scientific articles. *Scientometrics*, 111(3):1851–1855, 2017.

Sarah de Rijcke, Paul F Wouters, Alex D Rushforth, Thomas P Franssen, and Björn Hammarfelt. Evaluation practices and effects of indicator use—a literature review. *Research Evaluation*, 25(2):161–169, 2015.

Ed J Rinia, Th N Van Leeuwen, Hendrik G Van Vuren, and Anthony FJ Van Raan. Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the netherlands. *Research policy*, 27(1):95–107, 1998.

Arie Rip and J Courtial. Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6):381–400, 1984.

Anna Ritchie. Citation context analysis for information retrieval. Technical report, University of Cambridge, Computer Laboratory, 2009.

Mike Rossner, Heather Van Epps, and Emma Hill. Show me the data, 2007.

Ronald Rousseau. Sitations: an exploratory study. *Cybermetrics*, 1(1): 1–7, 2003.

Ronald Rousseau. Reflections on recent developments of the h-index and h-type indices. In *Proceedings of WIS 2008, Fourth International Conference on Webometrics, Informetrics and Scientometrics*, pages 1–8, Berlin, Germany, jun 2008. Humboldt-Universitat zu Berlin.

Ronald Rousseau et al. On the relation between the wos impact factor, the eigenfactor, the scimago journal rank, the article influence score and the journal h-index. Online preprint at `http://eprints.rclis.org/13304/`, 2009.

Royal Netherlands Academy of Arts and Sciences. Standard evaluation protocol 2009–2015: Protocol for research assessment in the netherlands. Technical report, Royal Netherlands Academy of Arts and Sciences (KNAW), 2009.

Royal Netherlands Academy of Arts and Sciences. Quality assessment of scientific research. `https://www.knaw.nl/en/topics/kwaliteit/quality-assessment-of-scientific-research/overview`, 2017. Accessed: 2017-07-15.

Javier Ruiz-Castillo and Ludo Waltman. Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, 9(1):102–117, 2015.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220.

Gerard Salton and Michael J. McGill. *Introduction to Modern Informa-*

tion Retrieval. McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN 0070544840.

Ian Sample. Harvard University says it can't afford journal publishers' prices. *The Guardian*, 04 2012. Accessed: 2017-05-15.

San Francisco DORA. San Francisco Declaration on Research Assessment: Putting science into the assessment of research. `http://www.ascb.org/files/SFDeclarationFINAL.pdf?x30490`, 2012. Accessed: 2017-07-01.

Christian Schlögl, Juan Gorraiz, Christian Gumpenberger, Kris Jack, and Peter Kraker. Are downloads and readership data a substitute for citations? the case of a scholarly journal. *Libraries in the digital age (LIDA) proceedings*, 13, 2014.

Science Europe. Science europe position statement on research information systems, 2016.

Science Watch. The Most-Cited Institutions Overall, 1999-2009. `http://archive.sciencewatch.com/inter/ins/09/09Top20Overall/`, 2009a. Accessed: 2016-04-30.

Science Watch. Top Ten Most-Cited Journals (All Fields), 1999-2009. `http://archive.sciencewatch.com/dr/sci/09/aug2-09_2/`, 2009b. Accessed: 2016-04-30.

SCImago. SJR – SCImago Journal & Country Rank. `http://www.scimagojr.com`, 2007. Accessed: 2016-04-29.

Scimago Lab. Scimago journal & country rank. `http://www.scimagojr.com/`, 2016. Accessed: 2017-04-23.

Scimago Lab. SJR SCImago Journal & Country Rank. `http://www.scimagojr.com/`, 2017. Accessed: 2017-09-09.

Per O Seglen. The skewness of science. *Journal of the American society for information science*, 43(9):628, 1992.

Per O Seglen. Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, 45 (1):1, 1994.

Per O Seglen. Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, 314(7079):498, 1997.

Xiaolin Shi, Jure Leskovec, and Daniel A McFarland. Citing for high impact. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 49–58. ACM, 2010.

Stuart Shieber. Why open access is better for scholarly societies. `https://blogs.harvard.edu/pamphlet/2013/01/29/why-open-access-is-better-for-scholarly-societies/`, 2013. Accessed: 2017-05-15.

Advaith Siddharthan and Simone Teufel. Whose idea was this, and why does it matter? attributing scientific work to citations. In *HLT-NAACL*, pages 316–323, 2007.

Thiago H. P. Silva, Mirella M. Moro, Ana Paula C. Silva, Wagner Meira Jr., and Alberto H. F. Laender. Community-based Endogamy as an Influence Indicator. In *Digital Libraries 2014 Proceedings*, page 10, 2014. ISBN 9781479955695.

Mikhail V Simkin and Vwani P Roychowdhury. Read before you cite! *arXiv preprint cond-mat/0212043*, 2002.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bojune Paul Hsu, and Kuansan Wang. An overview of microsoft academic

service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

Henry Small. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science*, 24(4):265–270, 1973.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine*, 99(4):178–182, 2006.

Harold C Sox and Drummond Rennie. Research misconduct, retraction, and cleansing the medical literature: lessons from the poehlman case. *Annals of Internal Medicine*, 144(8):609–613, 2006.

Padmini Srinivasan. Text mining: generating hypotheses from medline. *Journal of the Association for Information Science and Technology*, 55 (5):396–413, 2004.

Johannes Stegmann and Guenter Grohmann. Hypothesis generation guided by co-word clustering. *Scientometrics*, 56(1):111–135, 2003.

Robert J Sternberg and Tamara Gordeeva. The anatomy of impact: What makes an article influential? *Psychological Science*, 7(2):69–75, 1996.

Christian Sternitzke and Isumo Bergmann. Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, 78(1):113–130, 2009.

William J Sutherland, David Goulson, Simon G Potts, and Lynn V Dicks. Quantifying the impact and relevance of scientific research. *PLoS One*, 6(11):e27537, 2011.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009.

Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1293. ACM, 2012.

Dario Taraborelli. Soft peer review: Social software and distributed scientific evaluation. In *Proceedings of the 8th International Conference on the Design of Cooperative Systems (COOP '08)*, Carry-le-Rouet, France, may 2008.

Jaime A Teixeira da Silva and Judit Dobránszki. Problems with traditional science publishing and finding a wider niche for post-publication peer review. *Accountability in research*, 22(1):22–40, 2015.

Tertiary Education Commission. Performance-based research fund: Evaluating research excellence -– the 2012 assessment. final report. Technical report, Tertiary Education Commission, 2013a.

Tertiary Education Commission. Performance-based research fund: Quality evaluation guidelines 2012. Technical report, Tertiary Education Commission, 2013b.

Tertiary Education Commission. Performance-based research fund. http://www.tec.govt.nz/funding/funding-and-performance/funding/fund-finder/performance-based-research-fund/, 2017. Accessed: 2017-07-08.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110. Association for Computational Linguistics, 2006.

The PLoS Medicine Editors. The impact factor game. *PLoS medicine*, 3 (6), jun 2006.

The U.K. Cabinet Office. Realising our potential: A strategy for science, engineering and technology. Technical report, The U.K. Cabinet Office, 1993.

Mike Thelwall. Bibliometrics to Webometrics. *Journal of Information Science*, 34(4):1–18, 2007.

Mike Thelwall and Kayvan Kousha. Web indicators for research evaluation. part 1: Citations and links to academic articles from the web. *El profesional de la información*, 24(5), 2015a.

Mike Thelwall and Kayvan Kousha. Web indicators for research evaluation. part 2: Social media metrics. *El profesional de la información*, 24(5):607–620, 2015b.

Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R Sugimoto. Do altmetrics work? twitter and ten other social web services. *PloS one*, 8(5):e64841, 2013.

Thomson Reuters. Journal citation reports – journal source data. http://admin-apps.webofknowledge.com/JCR/help/h_

sourcedata.htm#sourcedata, 2012. Version: 2012-05-22, Accessed: 2017-01-26.

Times Higher Education. Citation averages, 2000-2010, by fields and years. https://www.timeshighereducation.com/news/citation-averages-2000-2010-by-fields-and-years/415643.article, 2011. Accessed: 2016-04-02.

Andrew Tomkins, Min Zhang, and William D Heavlin. Single versus double blind reviewing at wsdm 2017. *arXiv preprint arXiv:1702.00502*, 2017.

James R Troyer. In the beginning: the multiple discovery of the first hormone herbicides. *Weed Science*, 49(2):290–297, 2001.

William A Turner, G Chartron, F Laville, and B Michelet. Packaging information for peer review: new co-word analysis techniques, 1988.

UNESCO. Unesco institute for statistics. http://data.uis.unesco.org/, 2017. Accessed: 2017-04-22.

University of Washington. Eigenfactor.org faq. http://www.eigenfactor.org/faq.php, 2017. Accessed: 2017-09-09.

Marco Valenzuela, Vu Ha, and Oren Etzioni. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Guido Van Hooydonk. Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*, 48(10):944–945, 1997.

Richard van Noorden. Half of 2011 papers now free to read. *Nature*, 500: 386–387, 2013.

Anthony FJ Van Raan. Sleeping beauties in science. *Scientometrics*, 59 (3):467–472, 2004.

Alex D Wade, Kuansan Wang, Yizhou Sun, and Antonio Gulli. Wsdm cup 2016: Entity ranking challenge. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 593–594. ACM, 2016.

Richard Walker, Beatriz Barros, Ricardo Conejo, Konrad Neumann, and Martin Telefont. Personal attributes of authors and reviewers, social bias and the outcomes of peer review: a case study. *F1000Research*, 4, 2015.

Ludo Waltman. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391, 2016.

Ludo Waltman and Rodrigo Costas. F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, 65(3):433–445, 2014.

Ludo Waltman and Nees Jan van Eck. A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, 7(4):833–849, 2013.

Xiaojun Wan and Fang Liu. Are all literature citations equally important? automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, 65 (9):1929–1938, 2014.

Mingyang Wang, Guang Yu, and Daren Yu. Mining typical features for highly cited papers. *Scientometrics*, 87(3):695–706, 2011.

Xianwen Wang, Chen Liu, Wenli Mao, and Zhichao Fang. The open access advantage considering citation, article usage and social media attention. *Scientometrics*, 103(2):555–564, 2015.

Yujing Wang, Yunhai Tong, and Ming Zeng. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *AAAI*, 2013.

Zhong-Yi Wang, Gang Li, Chun-Ya Li, and Ang Li. Research on the semantic-based co-word analysis. *Scientometrics*, 90(3):855–875, 2012.

Andy R Weale, Mick Bailey, and Paul A Lear. The level of non-citation of articles within a journal as a measure of quality: a comparison to the impact factor. *BMC medical research methodology*, 4(1):14, 2004.

Marc Weeber, Henny Klein, Lolkje de Jong-van den Berg, Rein Vos, et al. Using concepts in literature-based discovery: Simulating swanson's raynaud–fish oil and migraine–magnesium discoveries. *Journal of the Association for Information Science and Technology*, 52(7):548–557, 2001.

Ian Wesley-Smith, Carl T Bergstrom, and Jevin D West. Static ranking of scholarly papers using article-level eigenfactor (alef). In *Proceedings of WSDM Cup 2016 – Entity Ranking Challenge at the 9th ACM International Conference on Web Search and Data Mining (WSDM 2016)*, 2016.

Ryan Whalen, Y Huang, A Sawant, B Uzzi, and Noshir Contractor. Natural language processing, article content & bibliometrics: Predicting high impact science. *ASCW*, 15:6–8, 2015.

Howard D. White and Belver C. Griffith. Author Cocitation: A Literat-

ure Measure of Intellectual Structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981.

White House Office of Science and Technology Policy. The 2014 budget: A world-leading commitment to science and research, 2014.

Robin Whitty. Some comments on multiple discovery in mathematics. *Journal of Humanistic Mathematics*, 7(1):172–188, 2017. doi: 10.5642/jhummath.201701.14.

James Wilsdon, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinnkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. *The metric tide: independent review of the role of metrics in research assessment and management.* Sage, 2015.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Jian Wu, Chen Liang, Huaiyu Yang, and C Lee Giles. Citeseerx data: semanticizing scholarly papers. In *Proceedings of the International Workshop on Semantic Big Data*, page 2. ACM, 2016.

Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M Chu, and Hongyuan Zha. On modeling and predicting individual paper citation count over time. In *IJCAI*, pages 2676–2682, 2016.

Jun Xu, Yaoyun Zhang, Yonghui Wu, Jingqi Wang, Xiao Dong, and Hua Xu. Citation sentiment analysis in clinical trial papers. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1334. American Medical Informatics Association, 2015.

Qinyi Xu, Andrea Boggio, and Andrea Ballabeni. Countries' biomedical publications and attraction scores. a pubmed-based assessment. *F1000Research*, 3, 2014.

Rui Yan, Congrui Huang, Jie Tang, Yan Zhang, and Xiaoming Li. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 51–60. ACM, 2012.

Zhenbin Yan, Qiang Wu, and Xingchen Li. Do hirsch-type indices behave the same in assessing single publications? an empirical study of 29 bibliometric indicators. *Scientometrics*, 109(3):1815–1833, 2016.

Guo Zhang, Ying Ding, and Staša Milojević. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the Association for Information Science and Technology*, 64(7):1490–1503, 2013.

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427, 2015.

George K. Zipf. *The psycho-biology of language: an introduction to dynamic philology.* The International Library of Psychology, London, 1935.

Part IV

Appendix

Appendix A

Survey on research publication quality

A.1 Email invitation

Subject line: What is research publication quality?

Dear <name>,

Please take a moment to consider this invitation to our survey. This is an opportunity for you to potentially influence the way research in UK higher education institutions is evaluated as the results of the survey will be provided to HEFCE, who jointly with SFC, HEFCW and DEL conducted the 2014 REF.

To access the survey, please go to <URL>.

Whether you will decide to take part in the survey or not, we would greatly appreciate your opinion on the matter!

What: Survey on academics' perception of research publication quality.

Why: To investigate the concept of research publication quality for use in research assessment.

How long: About 20 minutes.

Who: Drahomira Herrmannova and Petr Knoth, Knowledge Media Institute.

What we would like you to do: Answer a number of multiple-choice questions with your honest, personal opinion.

This survey is completely anonymous and personal details will be kept confidential.

Thank you in advance for your participation in our survey, we greatly appreciate the time you will take to complete it!

If you have any questions about the survey or our research, please contact research-quality-survey@open.ac.uk.

Sincerely,

Drahomira Herrmannova & Petr Knoth

A.2 Introduction

Survey title: Understanding research publication quality

The goal of this survey is to examine the concept of research publication quality for use in research assessment. This is an opportunity for you to potentially influence research evaluation in UK as the results of the survey will be provided to HEFCE, who jointly with SFC, HEFCW and

DEL conducted the 2014 REF.

The survey is aimed at understanding academics' perception of research publication quality. We aim to investigate the perceived differences between publication impact (in the traditional bibliometric sense), quality, rigour, significance and originality. We are also interested in aspects in which research quality is evidenced in publication manuscripts. We would greatly appreciate your opinion on the matter!

After answering few questions about your research expertise and experience, you will be presented with a list of statements for which you specify whether you agree or disagree. The survey also includes four open-ended questions. Answering the survey should take no more than 20 minutes.

This survey is completely anonymous and personal details will be kept confidential.

Thank you in advance for your participation in this survey. We greatly appreciate your time and effort!

Drahomira Herrmannova & Petr Knoth

A.3 Survey questions

Personal details

Explanatory text: Please provide the following information. All details will be kept confidential and anonymous.

Q: Which research area do you feel most associated with?

A: The list of disciplines presented to the respondents matched the units of assessment used in the latest Research Excellence Framework (REF) [Research Excellence Framework, 2014a].

Q: If you selected "Other" in the previous step, please provide details.

Q: Areas of interest. Provide a comma separated list of topics of your interest.

Q: Years since PhD or equivalent. If you don't have a PhD, please write "0".

Q: Number of authored publications.

A: Please select one of the following options.

- 5 or less
- 6-15
- 16-25
- 26-50
- 51-100
- More than 100

Examples of high quality research publications

Q: Please think of a few publications which you consider to be of very high quality. We would appreciate if you list them below, however this is not required for the survey. There are no requirements regarding the topic or the type of publication (e.g. primary research, survey, journal

article, conference item, etc.). You can list as many publications as you want, however please only provide publications in English.

Your perception of research publication quality

We would like to know your thoughts about research publication quality. Please provide answers to the following questions. A list of keywords or phrases as an answer to each of the questions is sufficient.

Q: Why do you consider the publications you listed in the previous step to be of high quality? Please provide a list of features which you think are an evidence of the quality of the selected publications.

Q: Is there something specific that you think is an important aspect of quality of research publications? What do you think makes a publication to be of high quality?

Q: How do you think the quality of research publications should be evaluated?

Aspects indicative of originality/novelty

We have listed a set of aspects which are demonstrative of research publications originality/novelty. We ask you to indicate how important is each of these aspects in your perception of originality. Please notice the scale is from 0 (not at all) to 10 (extremely) (there is a horizontal scroll bar below the list of aspects).

Q: In your perception, how indicative is each aspect of research publication originality/novelty? Please rate the following items on a scale from

0 to 10, 0 = not at all indicative, 10 = extremely indicative.

1. Provides new knowledge
2. Provides new data/resources enabling further research
3. Presents a new theory or theoretical framework
4. Presents a new method (methodology, experiment, test, technique, treatment, etc.)
5. Presents a new viewpoint on a problem
6. Clarifies existing problem(s)
7. Opens up a new problem (research question) for investigation
8. Provides new ideas
9. Provides evidence that supports an existing theory
10. Provides evidence that fails to support an existing theory
11. Integrates many different areas of data previously thought to be unrelated
12. Connects and integrates work from multiple disciplines
13. Integrates into a new, simpler framework data that had previously required a complex and possibly unwieldy framework
14. Contains generalisations, which are clearly stated, confirmed
15. Combining known methods in a new way
16. Applying known methods to a known problem for the first time

Q: If you think we have forgotten to mention any important aspects or for any other comments regarding the aspects of originality/novelty, please use the field below.

Aspects of rigour

We have listed a set of aspects which are demonstrative of research publications rigour. We ask you to indicate how important is each aspect in your perception of rigour. The scale is again from 0 (not at all) to 10 (extremely).

Q: Please indicate how important is each aspect in your perception of rigour. Please rate the following items on a scale from 0 to 10, 0 = not at all indicative, 10 = extremely indicative.

1. The problem is clearly stated and well-conceptualised
2. The publication presents the purpose and motivation for tackling the problem
3. The hypothesis is clearly stated
4. The publication uses a well-established methodology
5. The methodology selection matches the hypothesis and the data
6. If a new methodology is introduced, it is sound
7. If a new methodology is introduced, it is explained in enough detail
8. The publication contains a description of the data collection
9. The experiment is described in enough detail to be reproducible

10. The data used in the experiment are publicly shared and accessible
11. The data involve a sufficient number of cases (data, samples, events, patients etc.)
12. The results are checked for statistical significance
13. The results are valid
14. The publication presents valid but negative results
15. The results interpretation is unbiased and unambiguous
16. The results are discussed thoroughly (considering different interpretations and extreme cases)
17. The publication describes how the results were obtained
18. The publication discusses the contribution and importance of the results
19. The publication provides substantial and convincing evidence for proving or disproving the hypothesis
20. The publication objectively discusses the limitations of the results
21. The publication presents a proof of the results
22. The publication builds on previous research
23. The literature review section mentions all important relevant studies
24. Sources are cited for their importance and relevance (rather than collegiality, venue impact, etc.)
25. The literature review mentions in which way the paper makes a contribution to the field

26. Contains recommendations for further research
27. Contains implications for future research
28. Clear and concise abstract
29. Clear and concise conclusion
30. Consistent writing
31. Clear, concise and grammatically correct language
32. Unbiased tone
33. Keeping the writing to the point
34. Is easily understandable
35. The writing attracts and keeps attention
36. The paper is of an adequate length given the problem

Q: If you think we have forgotten to mention any important aspects or for any other comments regarding the aspects of rigour, please use the field below.

Aspects of significance

We have listed a set of aspects which are demonstrative of research publications significance. We ask you to indicate how important is each aspect in your perception of significance. The scale is again from 0 (not at all) to 10 (extremely).

Q: Please indicate how important is each aspect in your perception of significance. Please rate the following items on a scale from 0 to 10, 0 = not at all indicative, 10 = extremely indicative.

1. Topic is important
2. Topic is popular
3. Further research mentions the results
4. Further research builds on the results
5. Results encouraged a significant knowledge shift
6. Is criticised or scrutinised by further research
7. Influenced professional practice (policies, recommendations)
8. Is applicable in many areas
9. Influences multiple disciplines
10. Has resulted in a patent
11. Has resulted into a product or service
12. Has provided societal benefits (economic, social, etc.)
13. Has resulted in media coverage(e.g. news coverage, etc.)
14. Has generated public interest(e.g. as measured by tweets, non-academic invited talks, blog mentions, etc.)
15. Has received funding as a result of the research
16. Has been published in a high-impact journal
17. Has been presented at a high esteem conference
18. Has been publicly acknowledged by the research community
19. Has been read by a significant number of people(e.g. as measured by downloads, views, bookmarks, etc.)

20. Received citations from outside of its area/field
21. Received citations within its specialised area
22. Received many citations

Q: If you think we have forgotten to mention any important aspects or for any other comments regarding the aspects of significance, please use the field below.

Relation of originality, significance and rigour to quality

Q: How much do you agree with the following statements? Please indicate on a scale from 1 to 5 to what extent do you agree, with 1 = agree, 2 = somewhat agree, 3 = neither agree nor disagree, 4 = somewhat disagree, 5 = disagree.

1. How much do you agree with the following statements?
2. Publications providing novel/original ideas are of a higher quality.
3. A research publication lacking originality/novelty cannot be of a high quality.
4. High quality research publications present original/novel research.
5. The level of significance of a research publication is independent of its quality.
6. High-quality research publications have higher significance.

7. High significance of a research publication is an evidence of its quality.
8. Significant research publications are of high quality.
9. The quality of a research publication is independent of its rigour.
10. A low rigour research publication cannot be of high quality.
11. High rigour research publications are of high quality.
12. High quality research publications present rigorous research.

A.4 Survey end page

Thank you for your participation!

We very much appreciate your time.

If you know of other people that might be willing to participate in this survey, we would appreciate it if you would share with them a link to the survey.

Again, thank you very much for your help!

Q: Additional comments. If you have any additional comments on the topic of research publication quality or about the survey, please use the field below.

A.5 Results

The following section contains a complete list of statements shared by the respondents in their answers to the second open-ended question "Why do you consider the publications you listed in the previous step to be of high quality?" The statements shared here have been processed by splitting the answers into separate statements, merging similar statements, and grouping the statements into six categories (five categories one of which was split into two subcategories). The second column in each table shows how many times has each statement appeared in any of the answers.

Table A.1: Statements which were assigned to the category "originality".

#	**Statement**	**Count**
1	innovative	7
2	contribution to the field	4
3	ground breaking	4
4	new ideas	4
5	points research in new directions	3
6	new methods	3
7	novelty	2
8	solved outstanding problem	2
9	original	2
10	balanced/thorough literature review	1
11	unique literature review	1
12	opened path for research in the area	1
13	makes good points	1
14	novel techniques	1
15	first to answer a question	1
16	useful literature review	1
17	originality	1
18	extended reach of the field	1
19	clarifying insight	1

#	Statement	Count
20	clear contribution	1
21	original research	1
22	provides multiple interpretations	1
23	continues to contribute	1
24	offers something new	1
25	changes understanding of the field	1
26	significant contribution	1
27	useful answer	1
28	unusual answer	1
29	insights changed the field	1
30	first of its kind	1
31	links between theory and practice	1
32	novel solution	1
33	advances understanding	1
34	original thought	1
35	pushes the agenda	1
36	adds an interesting perspective	1
37	argues for the need to shift focus	1
38	informative	1
39	original contribution	1
40	challenges status quo	1
41	throught-provoking	1
42	innovative methodology	1
43	pushes boundaries	1
44	first to investigate a new topic	1
45	original ideas	1
46	paradigm shifting	1
47	new evidence	1
48	cointains good ideas	1
49	clarifies aspects of the field	1
50	new interpretations	1
51	inventive	1

#	Statement	Count
52	novel insights	1
53	new analysis	1
54	new results	1
55	produced sound knowledge	1
56	novel research problem	1
57	says something new	1
58	novel finding	1
59	changed direction of a field	1
60	fills gap in literature	1
61	useful conclusion	1
62	new information	1
63	new data	1
	total	85

Table A.2: Statements which were assigned to the category “rigour”.

#	Statement	Count
1	rigorous	4
2	comprehensive	3
3	data quality	2
4	evaluation	2
5	thorough	2
6	good/convincing evidence	2
7	methodological rigour	2
8	informed	2
9	sound technical/theoretical background	2
10	balanced/thorough literature review	2
11	extensive data	2
12	good analysis	2
13	many references	1
14	thorough evaluation	1
15	good scientific justification	1

#	Statement	Count
16	cutting-edge techniques	1
17	critical analysis of previous work	1
18	interesting methodology	1
19	thorough literature review	1
20	convincing results	1
21	non-trivial techniques	1
22	supported by data	1
23	analytical	1
24	constructive criticism	1
25	provides implications of findings	1
26	well performed	1
27	well thought out	1
28	deep	1
29	complete	1
30	good literature review	1
31	transparent methodology	1
32	expansive definitions	1
33	highly informed	1
34	illuminating analysis	1
35	critical approach	1
36	application of theory	1
37	thorough experiments	1
38	cutting-edge theory	1
39	cutting-edge method	1
40	accurate	1
41	solid conclusions	1
42	carefully done	1
43	good methods	1
	total	58

Table A.3: Statements which were assigned to the category "significance".

#	**Statement**	**Count**
1	essential reference in the field	3
2	influential	3
3	times read by respondent	2
4	reference for students	2
5	relevance to respondent's field	1
6	societal impact	1
7	times cited by respondent	1
8	widely used	1
9	essential reference	1
10	essential for respondent's research	1
11	impactful	1
12	relevant topic	1
13	significance of results	1
14	significance in the field	1
15	status in the field	1
16	important reference	1
17	addresses key issues	1
18	clinical outcome	1
19	relevant	1
20	applicable in practice	1
21	important	1
22	world leading	1
23	internationally competitive	1
24	useful	1
25	topic important to researchers and practitioners	1
26	relevant to professionals	1
27	useful to researchers	1
28	authoritative	1
30	impact on later research	1

#	Statement	Count
31	inspired subsequent research	1
32	used in teaching	1
33	significance in academia	1
34	significance in practice	1
35	relevant to important issue	1
36	says something that matters	1
37	highly visible	1
38	finding affected wider field	1
39	significant	1
40	relevant to a wide field of research	1
41	still relevant after a long time	1
42	high impact	1
43	conceptually important	1
	total	48

Table A.4: Statements which were assigned to the category "writing/ presentation".

#	Statement	Count
1	well written	7
2	clarity of presentation	6
3	well explained	2
4	clearly written	2
5	detailed	2
6	methodology explanation	1
7	long introduction	1
8	nice typesetting	1
9	coherence of presentation	1
10	intelligible interrogation of theory	1
11	clear to read	1
12	clear examples	1
13	no weak sections	1

#	Statement	Count
14	clear problem definition	1
15	easily understandable	1
16	apolitical	1
17	presents assumptions	1
18	readable	1
19	explicit research questions	1
20	choice of methodology explained	1
21	section on future work	1
22	useful summary	1
23	method secion	1
24	results section	1
25	thorough discussion	1
26	well structured	1
27	easy to read	1
28	accessibility	1
29	clearly argued	1
30	well argued	1
	total	44

Table A.5: Statements which were assigned to the category "external evidence".

#	Statement	Count
1	number of citations	9
2	peer review	4
3	journal impact factor	3
4	citations	2
5	peer reviewed journal	2
6	award	2
7	author prestige	2
8	journal prestige	2
9	cited by others	1

#	Statement	Count
10	venue acceptance rate	1
11	Nobel prize	1
12	journal publication	1
13	peer opinion	1
14	cited by prominent authors	1
15	cited by good papers	1
16	venue	1
17	quality of journal	1
18	citations from journal publications	1
19	recognition	1
20	wide circulation	1
21	publication venue	1
22	peer reviewed to high standard	1
23	author	1
24	robust peer review	1
25	vetted by internationally-based scholars	1
26	venue editor well known	1
27	venue publishes high quality research	1
	total	45

Table A.6: Statements which were assigned to the category "other".

#	Statement	Count
1	multi-disciplinary	3
2	high quality research	2
3	timeless	2
4	clear results	1
5	easy to reproduce	1
6	times cited by respondent	1
7	poses interesting research questions	1
8	addresses a well-established field	1
9	questions orthodoxy	1

#	Statement	Count
10	presents a whole idea	1
11	ideas of great depth	1
12	principled exposition	1
13	doesn't overstate achievement	1
14	beautifully constructed	1
15	inexhaustible message	1
16	quality of ideas	1
17	future focused	1
18	keeps giving in proportion to the effort expended reading	1
19	monolithic	1
20	results that are likely true	1
21	elegant results	1
22	difficult results	1
23	important techniques	1
24	excellent archival research	1
25	perceptive thinking	1
26	guidelines	1
27	good quality	1
28	interesting	1
29	varied chapters	1
30	intelectually challenging	1
31	cutting edge science	1
32	high standards	1
33	highly regarding	1
34	benchmark of research quality	1
35	qualitative research	1
36	up-to-date research	1
37	empirically interesting	1
38	theoretically interesting	1
39	findings that are likely true	1
40	quality	1
41	addresses small area	1

#	**Statement**	**Count**
42	enabled respondent to think better	1
43	original primary research	1
44	do not follow trends	1
	total	48

Appendix B

Collecting seminal publications and literature reviews

B.1 Email invitation

Subject line: Survey invitation – Collecting highly cited publications

Dear <name>,

Please take a moment to consider this invitation to our survey. The goal of the survey is to create a collection of highly cited publications from different areas of science. We are asking for your help because of your academic background.

This survey consists of two parts and should take just a few minutes to complete. The first part is aimed at understanding your research area and expertise. In the second part, we only ask you to list two publications from your research area: 1) a paper that represents a seminal work and

2) a paper that represents a survey of the field.

To access the survey, please go to <URL>.

This survey is completely anonymous and personal details will be kept confidential.

Thank you in advance for your participation in the survey, we greatly appreciate the time and effort you will take to complete the survey!

If you have any questions about the survey or our research, please contact research-quality-survey@open.ac.uk.

Sincerely,

Dasha Herrmannova

B.2 Introduction

Survey title: Collecting highly cited publications

The goal of this survey is to create a collection of highly cited publications from different areas of science. This survey consists of two parts and should take just a few minutes to complete. The first part is aimed at understanding your research area and expertise. In the second part, we only ask you to list two publications from your research area: 1) a paper that represents a seminal work and 2) a paper that represents a survey (review) of the field.

This survey is completely anonymous and personal details will be kept confidential.

Thank you in advance for your participation in this survey. We greatly appreciate your time and effort!

Dasha Herrmannova

B.3 Survey questions

Personal details

Explanatory text: Please provide the following information. All details will be kept confidential and anonymous.

Q: Which research area do you feel most associated with?

A: The list of disciplines presented to the respondents matched the units of assessment used in the latest Research Excellence Framework (REF) [Research Excellence Framework, 2014a].

Q: If you selected "Other" in the previous step, please provide details.

Q: Areas of interest. Provide a comma separated list of topics of your interest.

Q: Years since PhD or equivalent. If you don't have a PhD, please write "0".

Q: Number of authored publications.

A: Please select one of the following options.

- 5 or less

- 6-15
- 16-25
- 26-50
- 51-100
- More than 100

Examples of highly cited research publications

Explanatory text: Please think of two publications from your research discipline (these don't have to be your own publications), one representing a seminal work and one representing a survey of the area, and list these publications below. We would appreciate if you provide a DOI (Digital Object Identifier) or a URL for each of the publications, however title, authors and year of publication are also acceptable. Please only provide publications in English.

Q: **Seminal paper:** Please provide a DOI/URL or a title, a list of authors and a year of publication of a seminal paper from your research area.

Q: **Survey paper:** Please provide a DOI/URL or a title, a list of authors and a year of publication of a survey (review) paper from your research area.

Q: **Research area:** Please state which specific research area or topic do these two publications relate to.

B.4 Survey end page

We very much appreciate your time.

If you know of other people that might be willing to participate in this survey, we would appreciate it if you would share with them a link to the survey.

Again, thank you very much for your help!

Q: **Comments:** If you have any comments on the topic of bibliometrics/research evaluation or about the survey, please use the field below.

Appendix C

Do citations and readership identify seminal publications? Experiment results

C.1 Discipline-based model

Table C.1: Results of independent one-tailed t-test performed using citation and readership counts on all disciplines separately.

Discipline	p **(citations)**	p **(readership)**	**Total**
Geography, Environmental Studies and Archaeology	0.3404	0.2081	8
Biological Sciences	0.1748	0.4956	17
Computer Science and Informatics	0.0895	0.4517	43
Mathematical Sciences	0.2549	0.2518	14
Earth Systems and Environmental Sciences	0.1162	0.1645	18
Business and Management Studies	0.1191	0.1577	19
Physics	0.3819	0.1679	26
Education	0.1162	0.2146	26

Discipline	p **(citations)**	p **(readership)**	**Total**
Psychology, Psychiatry and Neuroscience	0.2443	0.2293	9
Politics and International Studies	0.2007	0.4275	6
Electrical and Electronic Engineering, Metallurgy and Materials	0.4260	0.3397	16
Sociology	0.4302	0.3955	7
Classics	0.1265	0.2113	4
Art and Design: History, Practice and Theory	0.2702	0.4565	5
Social Work and Social Policy	0.0910	0.3365	6
Economics and Econometrics	0.1525	0.3977	8
General Engineering	0.2079	0.1453	4
Anthropology and Development Studies	0.2920	0.2850	4
Aeronautical, Mechanical, Chemical and Manufacturing Engineering	0.2439	0.2015	4
Modern Languages and Linguistics	0.1557	0.1154	4
Public Health, Health Services and Primary Care	0.2056	0.1906	6
Total	-	-	254

The columns TN, TP, FN and FP in Tables C.2 and C.3 show the number of true negatives (papers correctly predicted as review), true positives (papers correctly predicted as seminal), false negatives (seminal papers incorrectly predicted as review), and false positives (review papers incorrectly predicted as seminal), respectively. The column "Opt." shows accuracy achieved with the optimal model, column t_{opt} shows the threshold identified by the optimal model, and column "Base." shows accuracy of the baseline model.

Table C.2: Classification results using citation counts as a feature, performed on all disciplines separately.

Discipline	**Acc.**	**Opt.**	**Base.**	t_{opt}	**TN**	**TP**	**FN**	**FP**	**Tot.**
Geography, Environmental Studies and Archaeology	0.38	0.75	0.50	41	2	1	3	2	8
Biological Sciences	0.29	0.65	0.53	50	4	1	7	5	17
Computer Science and Informatics	0.30	0.63	0.53	50	7	6	17	13	43
Mathematical Sciences	0.57	0.64	0.57	14	1	7	1	5	14
Earth Systems and Environmental Sciences	0.33	0.67	0.50	59	3	3	6	6	18
Business and Management Studies	0.47	0.68	0.53	197	6	3	6	4	19
Physics	0.62	0.62	0.50	916	12	4	9	1	26
Education	0.38	0.69	0.58	19	3	7	8	8	26
Psychology, Psychiatry and Neuroscience	0.44	0.67	0.56	31	1	3	2	3	9
Politics and International Studies	0.67	0.67	0.50	389	3	1	2	0	6

Discipline	Acc.	Opt.	Base.	t_{opt}	TN	TP	FN	FP	Tot.
Electrical and Electronic Engineering, Metallurgy and Materials	0.63	0.69	0.50	50	5	5	3	3	16
Sociology	0.71	0.86	0.57	2	2	3	1	1	7
Classics	0.75	1.00	0.50	25	2	1	1	0	4
Art and Design: History, Practice and Theory	0.20	0.60	0.60	0	0	1	2	2	5
Social Work and Social Policy	0.50	0.83	0.50	17	2	1	2	1	6
Economics and Econometrics	0.63	0.75	0.50	119	3	2	2	1	8
General Engineering	0.50	0.75	0.50	69	1	1	1	1	4
Anthropology and Development Studies	0.00	0.50	0.50	0	0	0	2	2	4
Aeronautical, Mechanical, Chemical and Manufacturing Engineering	0.75	0.75	0.50	2138	2	1	1	0	4
Modern Languages and Linguistics	0.75	1.00	0.50	38	2	1	1	0	4

Discipline	Acc.	Opt.	Base.	t_{opt}	TN	TP	FN	FP	Tot.
Public Health, Health Services and Primary Care	0.33	0.67	0.50	2	1	1	2	2	6
All	0.45	0.68	-	-	62	53	79	60	254

Table C.3: Classification results using Mendeley reader counts as a feature, performed on all disciplines separately.

Discipline	Acc.	Opt.	Base.	t_{opt}	TN	TP	FN	FP	Tot.
Geography, Environmental Studies and Archaeology	0.00	0.50	0.50	0	0	0	4	4	8
Biological Sciences	0.41	0.59	0.53	123	6	1	7	3	17
Computer Science and Informatics	0.40	0.53	0.53	0	0	17	6	20	43
Mathematical Sciences	0.07	0.57	0.57	0	0	1	7	6	14
Earth Systems and Environmental Sciences	0.78	0.78	0.50	96	5	9	0	4	18
Business and Management Studies	0.63	0.63	0.53	256	7	5	4	3	19
Physics	0.23	0.62	0.50	4	4	2	11	9	26
Education	0.62	0.62	0.58	1	4	12	3	7	26
Psychology, Psychiatry and Neuroscience	0.33	0.67	0.56	21	1	2	3	3	9

Discipline	Acc.	Opt.	Base.	t_{opt}	TN	TP	FN	FP	Tot.
Politics and International Studies	0.33	0.67	0.50	1	1	1	2	2	6
Electrical and Electronic Engineering, Metallurgy and Materials	0.50	0.63	0.50	43	7	1	7	1	16
Sociology	0.43	0.72	0.57	40	1	2	2	2	7
Classics	0.75	0.75	0.50	1	2	1	1	0	4
Art and Design: History, Practice and Theory	0.20	0.60	0.60	0	0	1	2	2	5
Social Work and Social Policy	0.17	0.50	0.50	0	0	1	2	3	6
Economics and Econometrics	0.50	0.62	0.50	77	3	1	3	1	8
General Engineering	0.50	1.00	0.50	82	1	1	1	1	4
Anthropology and Development Studies	0.75	0.75	0.50	15	1	2	0	1	4
Aeronautical, Mechanical, Chemical and Manufacturing Engineering	0.00	0.50	0.50	0	0	0	2	2	4
Modern Languages and Linguistics	0.50	1.00	0.50	36	1	1	1	1	4

Discipline	Acc.	Opt.	Base.	t_{opt}	TN	TP	FN	FP	Tot.
Public Health, Health Services and Primary Care	0.33	0.67	0.50	8	0	2	1	3	6
All	0.42	0.62	-	-	44	63	69	78	254

C.2 Year-based model

Table C.4: Results of independent one-tailed t-test performed using citation and readership counts on all publication years separately.

Year	p (citations)	p (readership)	Total
1999	0.3738	0.1951	8
2000	0.1706	0.0555	10
2001	0.1988	0.3102	15
2003	0.1096	0.3459	9
2004	0.4157	0.1629	10
2005	0.2115	0.3178	17
2006	0.3230	0.2259	14
2007	0.1570	0.1482	15
2008	0.2112	0.4029	14
2009	0.1199	0.0531	11
2010	0.1098	0.3501	21
2011	0.2064	0.2207	18
2012	0.1154	0.4622	17
2013	0.4370	0.1918	19
2014	0.2785	0.0731	13
2015	0.4661	0.1684	11
2016	0.0842	0.3098	17
Total	-	-	239

The columns TN, TP, FN and FP in Tables C.5 and C.6 show the number of true negatives (papers correctly predicted as review), true positives (papers correctly predicted as seminal), false negatives (seminal papers incorrectly predicted as review) and false positives (review papers incorrectly predicted as seminal), respectively. The column "Opt." shows accuracy achieved with the optimal model, column t_{opt} shows the threshold identified by the optimal model, and column "Base." shows accuracy of the baseline model.

Table C.5: Classification results using citation counts as a feature, performed on all years separately.

Year	**Acc.**	**Opt.**	**Base.**	t_{opt}	**TN**	**TP**	**FN**	**FP**	**Total**
1999	0.75	0.75	0.75	0	0	6	0	2	8
2000	0.60	0.70	0.70	0	0	6	1	3	10
2001	0.13	0.60	0.53	3	1	1	7	6	15
2003	0.67	0.89	0.56	374	3	3	2	1	9
2004	0.30	0.70	0.50	35	2	1	4	3	10
2005	0.47	0.59	0.59	472	8	0	7	2	17
2006	0.57	0.57	0.57	1559	7	1	5	1	14
2007	0.67	0.67	0.60	37	5	5	1	4	15
2008	0.43	0.71	0.50	197	2	4	3	5	14
2009	0.45	0.55	0.64	214	5	0	4	2	11
2010	0.62	0.71	0.57	1105	11	2	7	1	21
2011	0.50	0.67	0.56	59	3	6	4	5	18
2012	0.71	0.71	0.65	633	11	1	5	0	17
2013	0.63	0.79	0.79	240	12	0	4	3	19
2014	0.69	0.69	0.77	64	9	0	3	1	13
2015	0.64	0.73	0.73	96	7	0	3	1	11
2016	0.59	0.71	0.59	2	9	1	6	1	17
All	0.55	0.69	-	-	95	37	66	41	239

Table C.6: Classification results using reader counts as a feature, performed on all years separately.

Year	**Acc.**	**Opt.**	**Base.**	t_{opt}	**TN**	**TP**	**FN**	**FP**	**Total**
1999	0.50	0.75	0.75	0	0	4	2	2	8
2000	0.60	0.70	0.70	0	0	6	1	3	10
2001	0.53	0.67	0.53	57	3	5	3	4	15
2003	0.22	0.56	0.56	0	0	2	3	4	9
2004	0.60	0.60	0.50	15	3	3	2	2	10
2005	0.65	0.65	0.59	327	9	2	5	1	17
2006	0.21	0.57	0.57	39	3	0	6	5	14
2007	0.20	0.60	0.60	10	3	0	6	6	15
2008	0.50	0.57	0.50	2775	6	1	6	1	14
2009	0.45	0.55	0.64	382	5	0	4	2	11
2010	0.57	0.62	0.57	326	11	1	8	1	21
2011	0.39	0.61	0.56	1	2	5	5	6	18
2012	0.41	0.65	0.65	41	7	0	6	4	17
2013	0.79	0.84	0.79	823	14	1	3	1	19
2014	0.62	0.69	0.77	123	8	0	3	2	13
2015	0.73	0.82	0.73	1028	7	1	2	1	11
2016	0.59	0.65	0.59	35	9	1	6	1	17
All	0.51	0.65	-	-	90	32	71	46	239

Appendix D

Evaluating research with semantometrics – Experiment results

D.1 Results

Table D.1: Results of independent one-tailed t-test performed to test whether each feature helps to distinguish between seminal papers and literature reviews.

#	**Feature ID**	**Feature name**	p
1	S16	B sum	0.0000
2	S15	B range	0.0000
3	S13	B min	0.0000
4	S39	D range	0.0001
5	S37	D min	0.0001
6	S40	D sum	0.0004
7	S49	E min	0.0005
8	S51	E range	0.0008
9	B1	citations	0.0012

#	**Feature ID**	**Feature name**	p
10	S28	C sum	0.0022
11	S18	B stdev	0.0037
12	S31	C variance	0.0040
13	B4	S-RCR	0.0047
14	S20	B p25	0.0051
15	S19	B variance	0.0059
16	S30	C stdev	0.0066
17	S44	D p25	0.0066
18	B3	Citations per year	0.0073
19	S36	C kurtosis	0.0074
20	S42	D stdev	0.0077
21	S48	D kurtosis	0.0091
22	B2	Citations per author	0.0110
23	S6	A stdev	0.0115
24	S43	D variance	0.0119
25	S17	B mean	0.0139
26	S47	D skewness	0.0153
27	S7	A variance	0.0158
28	S21	B p50	0.0228
29	S41	D mean	0.0239
30	S32	C p25	0.0263
31	S5	A mean	0.0298
32	S52	E sum	0.0319
33	62	Mean author distance	0.0327
34	S8	A p25	0.0327
35	S35	C skewness	0.0353
36	S14	B max	0.0355
37	S29	C mean	0.0456
38	S11	A skewness	0.0472
39	S60	E kurtosis	0.0514
40	S38	D max	0.0536
41	S45	D p50	0.0630

#	**Feature ID**	**Feature name**	p
42	S59	E skewness	0.0759
43	S12	A kurtosis	0.0770
44	S33	C p50	0.1093
45	S55	E variance	0.1163
46	S9	A p50	0.1329
47	S54	E stdev	0.1573
48	S61	contribution	0.1890
49	S53	E mean	0.2129
50	S34	C p75	0.2438
51	A3	Altmetric score	0.2467
52	S3	A range	0.2721
53	S1	A min	0.2747
54	S22	B p75	0.2776
55	S57	E p50	0.2852
56	S2	A max	0.2886
57	S58	E p75	0.3030
58	F63	Author endogamy	0.3217
59	S56	E p25	0.3330
60	S24	B kurtosis	0.3407
61	S50	E max	0.3649
62	S26	C max	0.3689
63	S4	A sum	0.3859
64	S10	A p75	0.3939
65	A2	Readers' discipline count	0.3977
66	A1	Reader count	0.4431
67	S25	C min	0.4535
68	S46	D p75	0.4577
69	S23	B skewness	0.4775
70	S27	C range	0.4859

Table D.2: Classification performance when using individual features and all 203 publications. The features are listed in descending order of accuracy, which is shown in brackets.

#	**GNB**	**SVM**
0	B range (0.65)	B min (0.66)
1	B min (0.65)	B range (0.65)
2	D min (0.61)	D range (0.64)
3	C variance (0.60)	D min (0.64)
4	D range (0.59)	D kurtosis (0.63)
5	C p25 (0.59)	D skewness (0.62)
6	D skewness (0.59)	Citations (0.60)
7	C stdev (0.58)	C sum (0.59)
8	D kurtosis (0.58)	B p50 (0.59)
9	D p25 (0.58)	E min (0.58)
10	E min (0.58)	B mean (0.58)
11	A variance (0.58)	B p25 (0.58)
12	A stdev (0.58)	E range (0.58)
13	B p50 (0.58)	S-RCR (0.57)
14	E range (0.58)	C p25 (0.57)
15	B mean (0.57)	Citations per year (0.57)
16	B p25 (0.57)	Altmetric score (0.55)
17	C mean (0.57)	E sum (0.55)
18	D mean (0.57)	A p25 (0.55)
19	Citations (0.56)	A skewness (0.55)
20	C sum (0.56)	D p25 (0.55)
21	D variance (0.55)	C mean (0.54)
22	S-RCR (0.55)	Citations per author (0.54)
23	A mean (0.55)	C kurtosis (0.51)
24	B variance (0.55)	contribution (0.50)
25	Citations per author (0.55)	B max (0.50)
26	B stdev (0.54)	C skewness (0.48)
27	Altmetric score (0.54)	Readers count (0.47)

#	GNB	SVM
28	E sum (0.54)	Readers disciplines (0.44)
29	A p25 (0.53)	D mean (0.33)
30	C kurtosis (0.53)	C stdev (0.32)
31	D stdev (0.53)	B stdev (0.32)
32	Citations per year (0.53)	D stdev (0.31)
33	C skewness (0.52)	A stdev (0.21)
34	A skewness (0.51)	A mean (0.12)
35	B max (0.51)	C variance (0.08)
36	contribution (0.50)	A variance (0.07)
37	Readers count (0.49)	B variance (0.07)
38	Readers disciplines (0.47)	D variance (0.05)

Table D.3: Classification performance when using individual features and the subset of publications which contain additional author information. The features are listed in descending order of accuracy, which is shown in brackets.

#	GNB	SVM
0	B p25 (0.67)	B min (0.69)
1	B min (0.66)	B range (0.66)
2	D kurtosis (0.66)	D skewness (0.62)
3	B stdev (0.65)	D kurtosis (0.60)
4	B range (0.64)	B p25 (0.59)
5	D skewness (0.63)	D min (0.58)
6	B mean (0.63)	D range (0.58)
7	Author endogamy (0.61)	B p50 (0.57)
8	D mean (0.61)	S-RCR (0.57)
9	B variance (0.60)	A skewness (0.56)
10	A mean (0.60)	Author endogamy (0.55)
11	B p50 (0.59)	E min (0.54)
12	D variance (0.57)	D p25 (0.54)
13	D min (0.57)	Citations (0.53)

#	GNB	SVM
14	A p25 (0.57)	B mean (0.53)
15	D p25 (0.57)	E range (0.52)
16	E sum (0.56)	B variance (0.51)
17	Citations (0.56)	B max (0.51)
18	C stdev (0.56)	D variance (0.51)
19	S-RCR (0.56)	A stdev (0.51)
20	Altmetric score (0.56)	C stdev (0.51)
21	C variance (0.56)	Contribution (0.51)
22	C sum (0.56)	C variance (0.51)
23	A variance (0.56)	D stdev (0.51)
24	A skewness (0.55)	A mean (0.51)
25	D stdev (0.55)	A variance (0.51)
26	Citations per year (0.55)	Author distance (0.50)
27	A stdev (0.54)	Readers count (0.49)
28	E min (0.54)	PER auth (0.48)
29	E range (0.53)	B stdev (0.47)
30	D range (0.53)	D mean (0.46)
31	C mean (0.52)	C p25 (0.44)
32	Author distance (0.52)	A p25 (0.44)
33	C p25 (0.51)	C kurtosis (0.44)
34	B max (0.50)	Readers disciplines (0.43)
35	Readers count (0.50)	Citations per year (0.31)
36	C kurtosis (0.48)	Altmetric score (0.28)
37	Contribution (0.47)	C mean (0.27)
38	Readers disciplines (0.46)	C sum (0.25)
39	C skewness (0.41)	C skewness (0.19)
40	Citations per author (0.15)	E sum (0.10)

Table D.4: Feature importance obtained by training a gradient coosting classifier (GBC), and by recursive feature elimination (RFE) on all 203 publications. The features are listed in descending order of importance according to the two methods.

#	**GBC**	**RFE**
0	C sum	D min
1	D min	Readers count
2	B min	Citations/auth.
3	D kurtosis	C skewness
4	C variance	C kurtosis
5	Contribution	Contribution
6	C kurtosis	D kurtosis
7	B p50	C sum
8	A p25	E range
9	Readers count	B min
10	E min	PER year
11	D skewness	B range
12	B stdev	S-RCR
13	A skewness	A stdev
14	C skewness	D stdev
15	A mean	E sum
16	B mean	D p25
17	D variance	D mean
18	S-RCR	C stdev
19	B range	B mean
20	B variance	C p25
21	E sum	Altmetric score
22	E range	D skewness
23	B p25	A p25
24	D stdev	Readers disciplines
25	Citations/auth.	C mean

#	GBC	RFE
26	A variance	B p50
27	C mean	A skewness
28	B max	B variance
29	A stdev	E min
30	D p25	A mean
31	citations	B p25
32	Readers disciplines	Citations
33	C p25	B stdev
34	Citations/year	C variance
35	D range	D range
36	Altmetric score	B max
37	C stdev	A variance
38	D mean	D variance

Table D.5: Feature importance obtained by training a gradient coosting classifier (GBC), and by recursive feature elimination (RFE) on the subset of publications which contain additional author information. The features are listed in descending order of importance according to the two methods.

#	GBC	RFE
0	B min	D min
1	Contribution	D kurtosis
2	D kurtosis	C skewness
3	C kurtosis	A stdev
4	D min	Contribution
5	C variance	B min
6	B max	Author endogamy
7	A mean	B range
8	Author endogamy	C p25
9	A variance	C kurtosis
10	B stdev	C stdev

#	GBC	RFE
11	D stdev	D skewness
12	Readers disciplines	A skewness
13	D p25	Altmetric score
14	Author distance	S-RCR
15	Citations	Citations/auth.
16	B variance	D variance
17	D variance	Readers disciplines
18	S-RCR	Readers count
19	Citations/year	C mean
20	D skewness	D p25
21	D mean	B p25
22	Citations/auth.	E sum
23	C sum	B variance
24	B mean	Author distance
25	C mean	E min
26	Readers count	A mean
27	D range	D mean
28	Altmetric score	A p25
29	C p25	B mean
30	A stdev	Citations
31	A skewness	B max
32	E range	Citations/year
33	B range	E range
34	A p25	B stdev
35	B p50	B p50
36	E sum	D stdev
37	B p25	A variance
38	C skewness	C sum
39	E min	D range
40	C stdev	C variance